A computational model of stereoscopic 3D visual saliency

Junle Wang*, Matthieu Perreira Da Silva, *Member, IEEE*, Patrick Le Callet, *Member, IEEE*, and Vincent Ricordel, *Member, IEEE*

***Abstract*—Many computational models of visual attention performing well in predicting salient areas of 2D images have been proposed in the literature. The emerging applications of stereoscopic 3D display bring additional depth information affecting the human viewing behavior, and require extensions of the efforts made in 2D visual modeling. In this paper, we propose a new computational model of visual attention for stereoscopic 3D still image. Apart from detecting salient areas based on 2D visual features, the proposed model 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 then proposed and examined. For the performance evaluation of 3D visual attention models, we have created an eye-tracking database which contains stereoscopic images of natural content and is publicly available along with this paper. The proposed model gives a good performance, 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.**

***Index Terms*—Visual attention, 3DTV, saliency map, depth saliency, stereoscopy, eye-tracking**

I. Introduction

A. Background

The human visual system (HVS), receives a considerably large amount of information well beyond its capability to process all of it. To cope with large amounts of information, visual attention is one of the most important mechanisms deployed in the HVS to reduce the complexity of scene analysis [1]. Thanks to visual attention, viewers can selectively focus their attention on specific areas of interest in the scene. Two mechanisms of visual attention are usually distinguished: bottom-up and top-down [2]. Bottom-up attention is involuntary, signal driven, and independent of a particular viewing task; whereas top-down attention is voluntary and strongly dependent both on the viewing task and the semantic information. These two mechanisms interact with each other and affect the human visual behavior [3] [4] [5] [6] [7] [8].

The deployment of visual attention mechanisms in image processing systems has met with increasing interest in recent years. A variety of areas, including compression [9], retargeting [10], retrieval [11], and quality assessment [12] have been beneficial when provided with the information on the location that attracts the viewers' attention in a visual scene. To fully exploit the benefits of visual-attention-based processing systems, the regions of a scene that attract attention need to be computationally identified. This is why computational visual attention models are developed and implemented.

Note that the computational models of visual attention might focus on predicting sequences of gaze shifts and/or saliency maps. In this paper, 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".

In the literature, a great number of computational models of 2D visual attention have been investigated based on the visual features integration theory [13]. Most them are based on a bottom-up architecture, relying on a number of low-level features such as luminance, color, orientation, e.g. [14] [15] [16] [17]. Additionally, the concepts of rarity [18] or surprise [19] may be included. Due to the strong link between overt visual attention and eye movements [20], these models are typically validated by using a ground truth obtained by means of eye tracking experiments. The recorded eye movements can be post-processed and represented in two ways: successions of fixations and saccades; or a so-called fixation density map which identifies the ground truth locations of visual interest. When compared with the fixation density map, many computational models of visual attention have proven good at predict eye movements in the viewing of 2D images.

Nowadays, stereoscopic 3D content increases the sensation of presence through the enhancement of depth perception. For simplicity of notation, from now on, we will use the term 3D to refer to stereoscopic 3D in the remainder of this article. To achieve the enhancement of depth perception, binocular depth cues (such as binocular disparity) are introduced and merged together with other (monocular) depth cues in an adaptive way depending on the viewing space conditions. However, this change of depth perception also largely changes the human viewing behavior [21] [22]. Compared to the amount of studies on 2D images, only a small number of works related to 3D content visual attention can currently be found in the literature. Nevertheless, studies related to 3D visual attention have been recently gaining an increasing amount of attention because of the emergence of 3D content (in cinemas and at home) and the recent availability of high-definition 3D-capable acquisition and display equipment. The challenges and importance, as

Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet and Vincent Ricordel are with LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597 (Institut de Recherche en Communications et Cybernétique de Nantes), Polytech Nantes, France, (e-mail: wang.junle@gmail.com, {matthieu.perreiradasilva, patrick.lecallet, vincent.ricordel}@univ-nantes.fr, phone: +33-240483060).

well as several new applications of 3D visual attention were introduced by Huynh-Thu *et al.* in [23]. They described the conflicts met by the HVS while watching 3D-TV, how these conflicts might be limited and how visual comfort might be improved by knowing how visual attention is deployed. Several new application areas were also introduced, which can be beneficial when provided with the locations (including depth) of salient areas. These candidate applications exist in the different steps of a typical 3D-TV delivery chain, e.g. 3D video capture, 2D to 3D conversion, reframing and depth adaptation, and subtitling 3D movies. It is worthy to note that, in addition to 3D-TV, detecting the interesting parts of a 3D scene is also relevant to research topics in robotics.

The rising demand of visual-attention-based applications for 3D content increases the importance of computationally modeling 3D visual attention. However, two questions need to be addressed when developing a 3D visual attention model: (1) the influence of 2D visual features and (2) the influence of depth on visual attention deployment in 3D viewing condition. The first question concerns the possibility of adapting existing 2D visual attention models to 3D cases; the second question concerns the means by which depth information can be taken into account.

B. How the deployment of 3D visual attention is affected by various visual features: previous experimental studies

Based on observations from psychophysical experiments, several studies have started to examine both qualitatively and quantitatively how visual attention may be influenced by 2D visual features and additional binocular depth cues.

One of the early works, by Jansen *et al.* [24], investigated the influence of disparity on viewing behavior in the observation of 2D and 3D still images. They conducted a free-viewing task on the 2D and 3D versions of the same set of images. They found that the additional depth information led to an increased number of fixations, shorter and faster saccades, and broader spatial exploration. However, no significant difference was found between the viewing of 2D and 3D stimuli concerning the saliency of several 2D visual features including mean luminance, luminance contrast, and texture contrast. This consistence of the influence of 2D low-level visual features implied: (1) the importance of 2D visual feature detection in the design of a 3D visual attention model, and (2) the possibility of adapting existing 2D visual attention models for the modeling of 3D visual attention.

Liu *et al.* [25] examined visual features at fixated positions for stereo images with a natural content. Instead of comparing the viewing behaviors between 2D and 3D content viewing, they focused on comparing visual features extracted from fixations and random locations in the viewing of 3D still images. On the one hand, they demonstrated that the values of some 2D visual features, including luminance contrast and luminance gradient, were generally higher at fixated areas. On the other hand, their results also indicated that disparity contrast and disparity gradient of fixated locations were lower than those at randomly selected locations. These results are inconsistent with the results of Jansen *et al.* who found that observers consistently look more at depth discontinuities (high disparity contrast areas) than at planar surfaces. One limitation of Liu *et al.*'s study might lie on the quality of the ground truth disparity map. The disparity maps they used came from a simple correspondence algorithm rather than from depth range sensing systems or sophisticated depth estimation algorithms. The final results might thus have been affected by a considerable amount of noise in the estimated disparity maps.

Hakkinen *et al.* [21] examined the difference in eye movement patterns between the viewing of 2D and 3D versions of the same video content. They found that eye movements are more widely distributed for 3D content. Compared to the viewing of a 2D content, viewers did not only look at the main actors but also looked at some other targets on a typical movie content. Their result shows that depth information from the binocular depth cue provides viewers with additional information, and thus creates new salient areas in a scene. This result suggests the existence of a saliency map from depth, and a potential "summation" operation during the integration of 2D and depth saliency information. Conversely, Ramasamy *et al.*'s study [26], which bears on stereo-filmmaking, showed that the observers' gaze points could be more concentrated when viewing the 3D version of some content (e.g. the scenes containing long deep hallway).

Concerning the depth plane where fixations tend to be located, Wang *et al.* [27] examined a so-called 'depth-bias' in the task-free viewing of still stereoscopic synthetic stimuli. They found that objects closest to the observer always attract the most fixations. The number of fixations on each object decreases as the depth order of the object increases, except for the furthest object which receives a few more fixations than the one or two objects in front of it. The number of fixations on objects at different depth planes was also found to be time dependent. This result is consistent with the result of Jansen *et al.* [24]. Considering the influence of center-bias in 2D visual attention, these results indicate the existence of an additional location prior according to the depth in the viewing of 3D content. This location prior indicates the possibility of integrating depth information by means of a weighting.

Wismeijer *et al.* [28] examined if saccades were aligned either with individual depth cues or with a combination of depth cues, by presenting stimuli in which monocular perspective cues and binocular disparity cues conflicted. Their results indicate a weighted linear combination of cues when the conflicts are small, and a cue dominance when the conflicts are large. They also found that vergence is dominated only by binocular disparity. Their results imply that the interocular distance recorded by binocular eye-tracking experiment for 3D content should be compensated by taking into account the local disparity value.

C. Previous works on 3D visual attention modeling

As introduced in the previous sections, great efforts have been put into the study of the viewing behavior of 3D content. However, in terms of development of computational models, and compared to the body of 2D visual attention models,

TABLE I
MAIN FEATURES OF COMPUTATIONAL MODELS OF 3D VISUAL ATTENTION. NOTE THAT DW DENOTES DEPTH-WEIGHTING MODEL, DS DENOTES DEPTH-SALIENCY MODEL, AND SV DENOTES STEREO-VISION MODEL.

DW	**Depth information**	**Operation**	**Validation**
Maki *et al.* [29]	Relative depth	Assigned the target closer to observers with highest priority.	Qualitative assessment; no quantitative comparison to eye-tracking data.
Zhang *et al.* [30]	Perceived depth, pop-out effect	Irregular space conversion. Pixels closer to observers and in front of the screen are considered to be more salient.	Qualitative assessment; no quantitative comparison to eye-tracking data.
Chamaret *et al.* [31]	Relative depth	Weight each pixel in 2D saliency map by its depth value.	Qualitative assessment; no quantitative comparison to eye-tracking data.
DS	**Depth information**	**Operation**	**Validation**
Ouerhani and Hugli [32]	Absolute depth (distance), surface curvature, depth gradient	Extract depth features from depth map. Compute additional conspicuity maps based on depth features. Pool all the conspicuity maps (from 2D features and depth features).	Qualitative assessment; no quantitative comparison to eye-tracking data.
Potapova *et al.* [33]	Surface height, relative surface orientation, occluded edges.	Compute one saliency map for each (2D and depth) feature, then sum all the saliency maps.	Qualitative assessment and quantitative comparison to labeled ROIs.
SV	**Depth information**	**Operation**	**Validation**
Bruce and Tsotsos [34]	Disparity	Take two views as input. Add interpretive neuronal units for stereo-vision modeling into 2D computational model which use visual pyramid processing architecture.	Qualitative assessment; no quantitative comparison to eye-tracking data.

only a few computational models of 3D visual attention have been proposed. In the viewing of 3D content, experimental results have demonstrated strong influences of 2D visual features. However, due to the addition of new depth cues, depth features and their combination or conflicts [35] [36] with other monocular cues, a direct use of a 2D visual attention model for 3D content is neither biologically plausible nor effective. Furthermore, the disparity between two views can raise serious challenges on collecting 3D gaze points and on creating the fixation density maps used as ground-truth, since the gaze data needs to be extrapolated or processed to provide a notion of depth in relation with gaze direction or location [23].

In the literature, several computational models of 3D visual attention have been investigated. All of these models contain a stage in which 2D visual features are extracted and used to compute 2D saliency maps. However, these models can be classified into three different categories depending on the way they use depth information (see Table I for the main properties of models of each category):

- Depth-weighting models. This type of models (e.g. [29], [30] and [31]) does not contain any depth-map-based feature-extraction processes. Apart from detecting the salient areas by using 2D visual features, these models share a same step in which depth information is used as the weighting factor of the 2D saliency. The saliency of each location (e.g. pixel, target or depth plane) in the scene is directly related to its depth. Both 2D scene and depth map are taken as input. Note that the depth maps used in these models can be ground truth depth maps provided by depth detection equipment, or come from depth estimation algorithms using two or multiple views.
- Depth-saliency models. The models (e.g. [32] and [33]) in this category take depth saliency as additional information. This type of models relies on the existence of "depth saliency maps". Depth features are first extracted from the depth map to create additional feature maps, which are then used to generate the depth saliency maps. These depth saliency maps are finally combined with 2D saliency maps (e.g. from 2D visual attention models using color, orientation or intensity) by using a saliency map pooling strategy to obtain a final 3D saliency map. This type of model also takes the 2D scene and the depth map as input .
- Stereo-vision models. Instead of directly using a depth map, this type of models (e.g. [34]) takes into account the mechanisms of the stereoscopic perception in the HVS. Bruce and Tsotsos extend the 2D models that use a visual pyramid processing architecture [37] by adding neuronal units for modeling the stereo vision. Images from both views are taken as input, from which 2D visual features can be considered. In addition, the model takes into account the conflicts between two eyes resulting from occlusions or large disparities.

Most of the existing 3D visual attention models belong to the first and the second categories. Figure 1 summarizes the two different ways by which depth information is used in these two types of models. Both types of models have their respective advantages and limitations. The depth-weighting

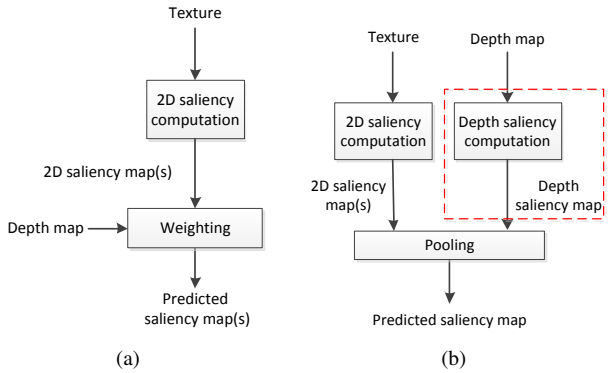


Fig. 1. Two different ways of using depth in (a) the depth-weighting models and (b) the depth-saliency models. Note that the main difference between these two types of models is the existence of a stage for extracting depth features and creating depth saliency map.

models can relatively easily adopt existing 2D models. The additional computational complexity is low due to the absence of depth feature extraction. However, a limitation of depth-weighting models is that they might fail to detect certain salient areas caused by depth features only. On the other hand, the depth-saliency models use depth as an additional visual dimension. They take into account the influence of depth features by creating depth saliency maps. This, however, the consideration of depth features increases the computational complexity. Besides, the influence of depth features on a model's performance has not been quantitatively validated.

D. Motivations

In the literature, most of the 3D visual attention models only take into account results of psychophysical experiments about depth's influence in a qualitative way. To our knowledge, any models that quantitatively integrates experimental observation results are still lacking. In terms of performance validation, eye-tracking data of 3D natural-content images containing various types of objects and scenes are crucial for evaluating the performance of computational models. However, this kind of database is still lacking. The absence of ground truth makes it difficult to quantitatively assess and compare the performances for most of the existing 3D computational models, and the influence of depth features as well. As a consequence, there is still not a strong agreement on how depth information should be used in 3D visual attention modeling: depth can be used to weight 2D saliency map; or alternatively it can be considered as an additional visual dimension to extract depth features leading to a depth saliency map.

In this paper, we propose a depth-saliency-based model of 3D visual attention. To benefit from psychophysical studies, we apply Bayes's theory on the result of an eye-tracking experiment using synthetic stimuli to model the correlation between depth features and the level of depth saliency. Concerning the integration of 2D saliency information, we propose to use a framework that can combine the resulting depth saliency map with existing 2D visual attention models, in order to exploit the state-of-the-art 2D models and achieve the prediction of the final 3D saliency map. Due to the lack of appropriate eye-tracking data for the performance evaluation, we have conducted a binocular eye-tracking experiment on 3D natural content images to create ground-truth. Given this ground truth, two methods to integrate depth information are also examined in this paper: a typical depth-weighting method and the propose depth saliency method.

Since a depth-saliency model highly relies on the extraction of depth features and the computation of a depth saliency map, we firstly introduce in section II a Bayesian approach of computing depth saliency map. A psychophysical experiment (the result of which is used for probability learning) is also introduced in this section. In section III, we introduce a framework of combining depth saliency map with 2D saliency map. Then, in section IV, we present a new eye-tracking database of 3D natural content images. In section V, the performance evaluation and a content-based analysis (on two ways to integrate depth) are carried out. The conclusion and discussion are presented in section VI.

II. Depth map and depth saliency map generation

A. Depth map creation

We propose that a depth map providing the information of the perceived depth of a scene needs to be computed at the first step of modeling 3D visual attention. In a stereoscopic 3D display system, depth information is usually represented by means of a disparity map which shows the parallax of each pixel between the left-view image and the right-view image. In the literature, a disparity map is usually directly adopted as depth information [31]. However, we propose too add a transformation from a disparity map to a depth map, which represents perceived depth in unit of length, in the chain of 3D visual attention modeling, since even the same disparity value corresponds to different perceived depth depending on the viewing condition.

Disparity is measured in unit of pixels for display systems. The relationship between disparity (in pixel) and perceived depth can be modeled by the following equation:

$$D = V/(1 + \frac{I \cdot W}{P \cdot R_x}) \quad (1)$$

where D represents the perceived depth, V represents the viewing distance between observer and screen plane, I represents the interocular distance, P is the disparity in pixels, W and R_x represent the width (in cm) and the horizontal resolution of the screen, respectively.

According to Equation 1, the perceived depth is not only a function of disparity but is also influenced by the viewing condition, which concerns the viewing distance and the properties of the display. For instance, an 8-pixel negative disparity can create a perceived depth of about 3.5 cm behind the screen when it is presented on a 24-inch full-HD stereoscopic display with a 1-meter viewing distance (3 times the height of the screen). However, the same disparity corresponds to a perceived depth of the infinite on an 8-meter high 2k cinema screen with an 8-meter viewing distance. When the viewing condition varies, the change of the perceived depth from even a same disparity value might make some areas of a 3D scene difficult to be fused. Consequently, the saliency distribution

can be different. In this paper, we adopt Equation 1 to compute the depth map for each image, the interocular distance is set to 6.3 cm, while the screen property parameters are set according to the setup of the eye-tracking experiment (introduced in section IV).

This stage of creating depth map can be beneficial for both the depth-weighting models and the depth-saliency models. For the depth-weighting models, the resulting depth map can be directly adopted as depth information for the weighting or the depth-based pooling of saliency maps. However, for the depth-saliency models, some more computations are necessary. The resulting depth map is taken as input for generating the depth saliency map.

B. A Bayesian approach of depth saliency map generation

In the area of saliency map creation, Bayes's theorem has been widely applied in various ways (e.g. [18], [38] and [39]). In this paper, we propose a new approach to apply Bayes's theorem for computing depth saliency maps based on features extracted from a depth map. The proposed approach correlates depth features with the degree of depth saliency, by using the data from a psychophysical experiment.

We first introduce the proposed definition of depth saliency: the depth saliency (S) of each location (a pixel) equals the probability of this point being gazed at, given the depth features observed from this point:

$$S = P(C = 1|\bar{f}_{dep}) \tag{2}$$

where C is a binary random variable denoting whether or not a point is gazed at. The random variable vector $\bar{f}_{dep}$ denotes depth features observed from this point. Note that the term about 'features', $\bar{f}_{dep}$, can stand not only for the local features (e.g. the distance to observer), but also for some higher order features considering the information from the neighborhood, such as the result of applying Difference of Gaussian kernel (DoG) on feature maps. By using Bayes' rule, we can obtain:

$$P(C = 1|\bar{f}_{dep}) = P(C = 1) \cdot \frac{P(\bar{f}_{dep}|C = 1)}{P(\bar{f}_{dep})} \tag{3}$$

Equation 3 represents how the depth features observed from a given point influence the probability of the HVS to decide whether to fixate this point or not. Here we make an assumption that, without any given features, the probability of a pixel to be fixated (i.e. $P(C = 1)$) is simply a constant. Therefore, the probability of each pixel to be fixated is proportional to the feature distribution at gazed points (i.e. $P(\bar{f}_{dep}|C = 1)$), normalized by the rarity of features in the context (i.e. $P(\bar{f}_{dep})$). Note that the use of the likelihood, $P(\bar{f}_{dep}|C = 1)$, in the proposed approach differs from the ways in which it is usually used by many models in the literature also applying Bayes's theorem. We do not do any binary classification to decide whether a point is a fixation or not. Instead, we define the result, the depth saliency map, as a distribution of probability of the points being gazed at as a function of depth features.

To achieve the computation of a depth saliency map, the proposed approach consists of two stages: depth feature extraction, and probability distribution modeling.

1) Depth feature extraction: The proposed definition of saliency can take into account various depth features. Nevertheless, in this paper, we particularly focus on using only depth contrast as the feature for depth saliency map prediction. Using fewer features decreases the computational complexity of the model. Note that depth contrast has been demonstrated to be a dominant feature in depth perception [40]. It is believed that depth is perceived most effectively at surface discontinuities [41]. In most situations, depth contrast can also be an efficient indicator of an interesting target. For example, the HVS might consider a region protruding above a flat plane as a potential target [33]; or might consider a hole as a place where a potential target might exist. In our study, Difference of Gaussians (DoG) filter is applied to the depth map for extracting depth contrast. We use DoG filter since it has been widely used by visual saliency models in the literature due to the resemblance to the receptive fields of neurons and the capability to simulate the center-surround mechanism in the HVS. The DoG filters used in the proposed model were generated by:

$$f(x,y) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) - \frac{1}{2\pi K^2\sigma^2}\exp(-\frac{x^2+y^2}{2K^2\sigma^2}) \tag{4}$$

where (x, y) is the location in the filter. σ and K were used to control the scales of DoG and the ratio between the 'center' area and 'surround' area. Since we do not find any mention of mechanisms related to multi-scale depth perception, we apply only one scale of DoG for a higher efficiency. In this paper, we selected a scale as $\sigma = 32$ pixels (approximately corresponding to 1 degree of visual angle in our experiment) and a center/surround ratio (i.e. 1/K) as 1/1.6 (the same value as the one used in [18]).

2) Probability distribution modeling: The function $P(C = 1|f_{contrast})$ models the relationship between the depth contrast of each position and the probability that this position is gazed at. We propose to model this function using a probability-learning of eye movement data collected from a free-viewing eye-tracking experiment.

An important factor that affects the modeling is the stimuli used in the eye-tracking experiment. We prefer to use synthetic stimuli rather than natural content stimuli. Generally, 3D images of natural content contain not only depth but also many other features affecting the eye movements. For instance, observers' attention could be affected by 2D bottom-up visual features such as color and luminance; or top-down features such as the presence of people, animals, or text; or center-bias caused by the preference of photographers to put the interesting objects close to the center of the scene. The simultaneous appearance of multiple features increases the difficulty of evaluating how people's viewing behavior is actually affected by depth information. On the other hand, obtaining a precise depth map for natural content 3D images is still challenging in terms of costs as well as of the quality of the depth map.

In our study, synthetic stimuli were used for the eye-tracking

experiment. These stimuli consisted of 3D scenes in which a background and some identical objects were deliberately displayed at different depth plane.

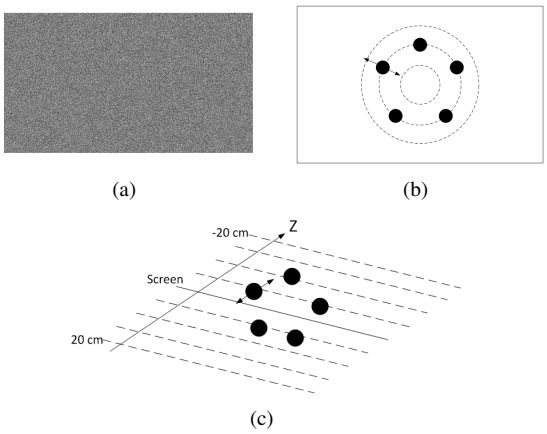


Fig. 2. Synthetic stimulus generation. (a) Background. (b) Projections of objects on the screen plane. (3)Arrangement of objects in depth.

The background was a flat image consisting of white noise as shown in Figure 2(a), which was placed at a depth value of -20 cm (20 cm beyond the screen plane). In each scene, the objects consisted of a set of black disks of the same diameter S. They were displayed at different depth values randomly chosen among $\{-20, -15, -10, -5, 0, 5, 10, 15, 20\}cm$ (see Figure 2(c)). The projections of the objects on the screen plane are uniformly positioned on a circle centered on the screen center (Figure 2(b)). Thus, it can be assumed that no "center-bias" was introduced in the observation. Note that all the objects and the background are within the comfortable viewing zone [42] considering the viewing conditions.

Concerning the objects, three parameters of the objects were varied from one scene to another: number, size, and distance from the screen center. The range of each parameters is introduced in [27]. Based on the combinations of the parameters, we created a set of 3186 scenes. One hundred and eighteen scenes were presented to each observer. Each scene was presented for 3 seconds. Twenty-seven subjects, ranging in age from 18 to 44 years, participated in the eye-tracking experiment. The details about the apparatus used for the experiment as well as the tests of a subject's visual acuity and 3D acuity are presented in section IV.

There are several advantages to using the proposed synthetic stimuli to collect training data for learning the relationship between depth features and people's viewing behavior. (1) It is possible to precisely control the depth of the objects and of the background. In other words, a precise depth map can be created for each scene. Moreover, due to the lower cost of generating synthetic stimuli, a great amount of stimuli can be exploited. (2) The influence of 2D visual features on viewing behavior can be limited. In our experiment, all the objects were uniformly located, with a constant shape, size, and distance from the center of the screen. This setup enables the stimuli to eliminate as many bottom-up visual attention features as possible. (3) The influence of depth features coming from depth cues other than disparity can be limited. Disparity was the only depth cue elicited in this experiment. The reason for choosing binocular disparity is that its relationship with the perceived depth can be easily modeled (as introduced in section II-A). For other (monocular) depth cues (e.g. perspective, occlusion or blur [43]), the influence on the perceived depth is difficult to be quantitatively measured. (4) The low complexity of the scenes allowed a shorter observation duration. The viewing time of natural content images in eye-tracking experiments was generally set to 5 seconds or more. The viewing time in our experiment was relatively shorter (3 seconds for each condition). Nevertheless, it was still long enough for the participants to subconsciously position their fixations on objects and explore the scene as they wanted. Hence, these simple stimuli enabled experimenters to collect more data.

The probability distribution $P(f_{contrast})$ can be obtained based on the depth contrast maps of the synthetic stimuli. By considering the probability distribution of depth contrast at gaze points recorded during the viewing, $P(f_{contrast}|C = 1)$ can be then obtained. Therefore, the likelihood $P(C = 1|f_{contrast})$ which models the relationship between the depth contrast and the probability of being fixated can be obtained by Equation 3. In Figure 3, we illustrate the resulting likelihood distribution $P(C = 1|\bar{f}_{dep})$. As seen in the figure, the saliency is not symmetrically distributed for positive and negative depth contrast values. For the positive values which correspond to protruding regions, the curve appears to be a linearly increasing line. A higher positive contrast value yields a larger chance on a fixation. In our experiment, higher positive contrast values result from larger distance between the objects and the background. For the negative feature values, which correspond to the 'dents', the curve is similar to a logarithmic curve: as the absolute of the contrast value increases, the chance on a fixation also increases, but at a slower rate. Both parts of the curve show that the depth saliency is highly related to depth contrast. Meanwhile, the asymmetry of the curve implies that the protruding objects are more likely to be gaze at.

For the implementation of the proposed model, the modeled $P(C = 1|f_{contrast})$ is applied on the depth feature map. By taking the depth contrast value at each pixel as input, the saliency value of each pixel in an image can be thus computed.

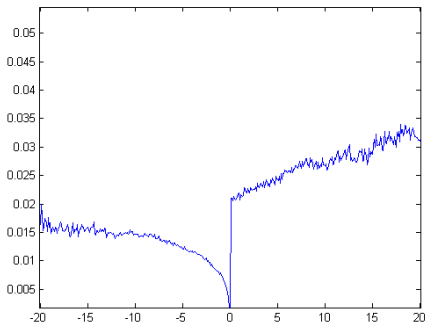


Fig. 3. The distribution $P(C = 1|\bar{f}_{contrast})$ resulting from the eye-tracking experiment using synthetic stimuli.

III. A Framework of computational model of 3D visual attention based on depth saliency

In this section, we introduce the framework which integrates the depth saliency map with the saliency maps computed from 2D visual features, and realizes the prediction of the final 3D saliency map. The general architecture of the proposed framework is presented in Figure 4.

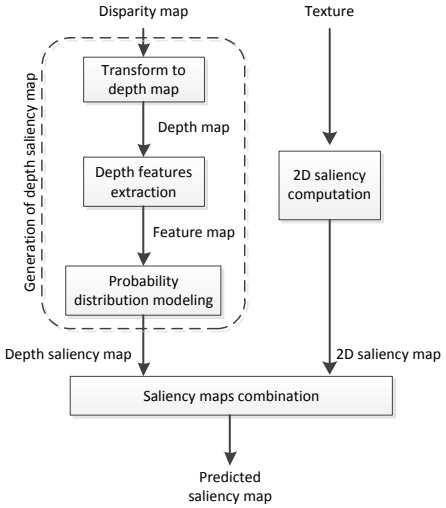


Fig. 4. Overview diagram of the proposed model

A. 2D saliency map generation

Since developing a completely new computational model of 2D visual attention is not in the scope of the present paper, we leave the work of 2D saliency map creation to existing models. Three bottom-up visual attention models using quite different mechanisms were used to perform the 2D saliency prediction, and involved in the final performance evaluation:

- Itti's model [14] performs a hierarchical decomposition based on three low-level visual features: luminance, color and orientation. The Matlab source code, saliencytoolbox [44], we used in this paper can be downloaded from the page: http://www.saliencytoolbox.net/. We obtained the saliency maps by performing the 'batchSaliency' command with default parameters.
- AIM model from Bruce [15] is based on a premise that localized saliency computation serves to maximize information sampled from one's environment. The source code used can be downloaded from the page: http://www-sop.inria.fr/members/Neil.Bruce/. We used the default parameters except that the rescaling factor was set to 0.25 (which means the input image was rescaled to 1/4 of its original size before the processing) to speed up the computation.
- Hou's model [16] computes the Fourier spectrum based on luminance only, and analyzes the spectral residual of an image. The source code used can be downloaded from the page: http://www.klab.caltech.edu/ xhou/. We used the default parameters.

In the proposed model, 2D saliency computation is only performed based on the image from the left view which is selected arbitrarily, since the images from the two views are quite similar, and the difference in 2D features between the images of the two views has thus only a marginal influence on visual attention deployment. Computing a 2D saliency map based on only one of the views instead of both views can help reduce the computational complexity.

B. Saliency maps combination

The goal of this saliency maps combination stage is to mix together the saliency maps obtained from different visual dimensions (i.e. depth information and 2D visual features in this paper). Since the 2D saliency map input is already the result of a pooling stage contained in the applied 2D visual attention model, this saliency maps combination stage focuses on merging only one 2D saliency map with the depth saliency map.

In the literature, although several approaches combining conspicuity maps of 2D visual features have been proposed, any specific and standardized approaches are still lacking to combine saliency maps from depth with 2D visual features. In the proposed model, we adopt a straightforward approach which is the same as the one used in [33] to merge the depth saliency map (SM_{dep}) with the 2D saliency map (SM_{2D}): the final saliency map SM_S is equal to the sum of both maps:

$$SM_S(i,j) = \omega_1 SM_{dep} + \omega_2 SM_{2D} \quad (5)$$

where $\omega_1 = \omega_2 = 0.5$.

IV. Eye-tracking database

So far, the lack of ground truth has limited the studies of computational models of 3D visual attention. To evaluate the performance of computational models, we create and publish a new eye-tracking database containing eighteen stereoscopic natural content images, the corresponding disparity maps, and the eye movement data for both eyes. This database [45] can be downloaded from the page: http://www.irccyn.ec-nantes.fr/spip.php?article1102.

A. Stimuli

The stereoscopic images used in the proposed database were acquired from two sources: (1) the Middlebury 2005/2006 image dataset, and (2) the IVC 3D image dataset.

1) The Middlebury 2005/2006 dataset: Scharstein *et al.* [46] created 30 multi-view 3D images. Each image corresponds to one particular indoor scene taken from a close-up view. Each of them consists of 7 rectified views. In this image acquisition system, the focal length was set to 3740 pixels, and the directions of the cameras were parallel. The ground-truth disparity maps were created by using an automated version of the structured-lighting technique of [47]. The images have a resolution about 1300*1100 pixels, and about 150 different integer disparity values.

We selected 10 images from the Middlebury 2005/2006 image dataset for our eye-tracking experiment (see Figure 5). Considering the visual comfort, the view 2 and the view 4 were used as the left view and the right view respectively for each scene. This selection was made to avoid the appearance of excessive relative disparity in one scene. The baseline between the two views was thus supposed to be 800 mm.

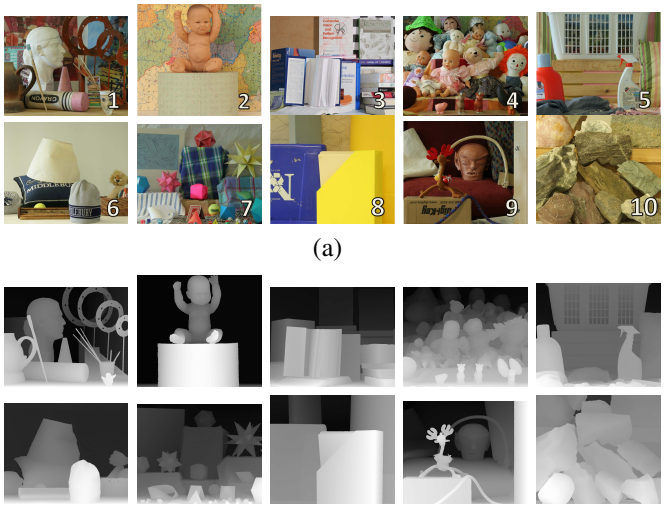


Fig. 5. (a) Images from the Middlebury image dataset. The number in the bottom right corner indicates the index of each image. (b) The corresponding disparity maps.

a) Stereo window violation removal: Since the cameras used for generating the views were set in a parallel direction, they are assumed to converge at an infinite point. This setup makes a direct utilization of the view 2 and the view 4 as a stereo pair lead to a so-called "stereo window violation" [48]. All the pixels in the scene were perceived as being in front of the screen plane; meanwhile, certain areas close to the left edge of the left view and certain areas close to right edge of the right view were displayed only in the left view and right view, respectively. Serious visual rivalry and visual discomfort can happen when these two areas are looked at.

Apart from visual rivalry, another problem is an insufficient exploitation of the depth range. [42] suggested that a 3D scene should be located in a limited depth range named comfortable viewing zone ranging from the back to the front of the screen plane. In our case, if the entire scene was displayed only in front of the screen plane, the depth range that could be exploited would thus be limited to approximately its half.

To overcome these problems, we adjusted the depth range of the scene by using the method proposed by Chamaret *et al.* [31]. We shifted the left view to left, the right view to the right. This shifting of the two views in opposite directions equals to adding a constant negative disparity for every pixel in the two views. The amount of added disparity was calculated by:

$$D_{add} = \frac{D_{\min} - D_{\max}}{2} \tag{6}$$

where $D_{\min}$ and $D_{\max}$ denote the minimum and maximum disparity values in the scene, respectively. Consequently, half of the depth range of the scene was moved to the back of the screen plane, while the other half was still in front of the screen plane.

b) Disparity map refining: Although most of the areas in the disparity maps of the images provided by the Middlebury dataset were with high accuracy, the disparity values were still unknown at some locations, such as some deep holes surrounded by several objects and the edges where occlusion happened (as shown in Figure 6).

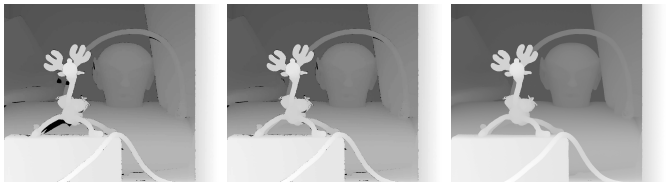


Fig. 6. Example of disparity map refining.

In the first case, the area of these regions was usually large, and the actual disparity value was different from all the surroundings. We thus did a manual refining by justifying the depth value of these regions considering the content of the whole scene. We first checked whether a region was part of the background or of any object, both of which have reliable depth information at some other locations in the scene, then we manually assigned the same depth value to these regions.

In the second case, the region with unknown disparity value in the disparity map usually consisted of some (groups of) pixels which covered small areas and were sparsely located along the edges. The disparity values of these pixels were close or even equal to the surrounding pixels. An automatic refining was thus performed by using an inpainting algorithm proposed by Criminisi *et al.* [49], which was an exemplar-based inpainting algorithm which fills holes in a visually plausible way and persists one-dimensional patterns, such as lines and object contours.

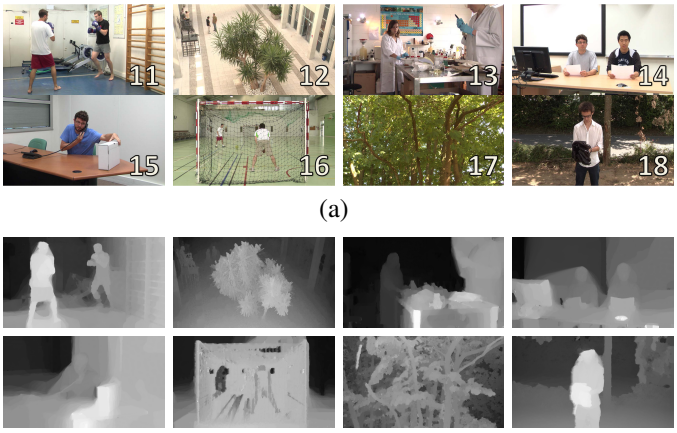


Fig. 7. (a) Images from the IVC 3D image dataset. The number in the bottom right corner indicates the index of each image. (b) The corresponding disparity maps.

2) The IVC 3D image dataset: We produced a set of eight 3D videos by using a Panasonic AG-3DA1 3D camera [50]. One frame from each video was selected by the authors to create this IVC 3D Image Dataset. Each video consists of two sequences representing the left and the right views, respectively. Both sequences were in full-HD resolution (1920*1080 pixels). This set of videos contains two outdoor scenes and six indoor scenes, which were taken in University of Nantes. Compared to the Middlebury database, the scenes in this set of videos have a higher average depth value. The distance

between the camera and the first object in the scene is at least two meters.

Without the use of any depth range sensors during the acquisition of videos, the depth maps of the IVC 3D image database were obtained by a post-processing depth map estimation on the stereo-pair images. The depth map estimation we applied was an optical flow approach proposed by Werlberger *et al.* [51] [52]. The general idea of this approach was inspired by 2D motion estimation algorithms that use optical flow estimation. To create the ground-truth disparity map, we computed the 'left-to-right' disparity map which represents the displacement of each pixel in the left view. Both the images and their corresponding disparity maps are showed in Figure 7.

B. Apparatus and procedures

Stimuli were displayed on a 26-inch (552×323 mm) Panasonic BT-3DL2550 LCD screen, which has a resolution of 1920×1200 pixels, and the refresh rate was 60 Hz. Each screen pixel subtended 61.99 arcsec at a 93 cm viewing distance. The maximum luminance of the display was 180 cd/m^2, which yielded a maximum luminance of about 60 cd/m^2 when watched through glasses. Observers viewed the stereoscopic stimuli through a pair of passive polarized glasses at a distance of 93 cm. The environment luminance was adjusted according to each observer, so that the pupil had an appropriate size for eye-tracking. SMI RED 500 remote eye-tracker was used to record the eye movements. A chin-rest was used to stabilize the observer's head.

The eighteen scenes were presented in a random order. The presentation time of each scene was 15 seconds. Between every two scenes, a center point was showed for 500 ms at the screen center with zero disparity. Note that the 15-second presentation time is relatively long as compared to the eye-tracking for 2D images. A discussion on the effect of the different presentation times and their impacts on performance evaluation for saliency models is presented in Section VI-B. In our experiment, subjects were required to do a free-viewing of the scene. A nine-point calibration was performed at the beginning of the experiment, and repeated every ten scenes. Note that the calibration points were displayed on the screen plane. The quality of the calibration was verified by the experimenter on another monitor. Participants could ask for a rest before each calibration started. Each observer was required to have at least three rests during the whole observation. All the experiments were conducted from 10:00 to 12:00 a.m. and from 2:00 to 4:00 p.m., in order to limit the feeling of fatigue as much as possible.

C. Participants

Thirty-five subjects participated in the experiment. Note that none of the subjects in this group had ever participated in the experiment for probability distribution modeling. Subjects ranged in age from 18 to 46 years old. The mean age of subjects was 24.2 years old. All the subjects had either normal or corrected-to-normal visual acuity, which was verified by three pretests prior to the start of the eye-tracking experiment: the Monoyer chart was used to check the acuity (subject must obtain results higher than 9/10); the Ishihara test was used to check color vision (subject should have no color troubles); and the Randot stereo test was used to check the 3D acuity (subject should get results higher than 7/10). All the subjects were also naive to the purpose of the experiment, and were compensated for their participation.

D. Fixation density map creation

In order to take into account both the position and duration of the eye movements, all gaze points recorded by the eye-tracker from both the left and the right eyes were used to create the fixation density maps. The gaze points maps from each eye were first created respectively. The left gaze points map was created by directly using the coordinates of the gaze positions of the left eye. However, according to the argument that it would be compelling, in a biological sense, to accommodate shifts in the position of an attended event from one eye to another [23], we created the right gaze points map by adding a displacement, horizontally and vertically, on the coordinates of each right-eye gaze point. The displacements of each gazed point were obtained from the 'right-to-left' disparity map computed with the same approach as the one used to create the ground-truth disparity maps of the IVC 3D images database.

The two gaze points maps were then summed and filtered by a two-dimensional Gaussian kernel to account for 1) the decrease in the visual accuracy with increasing eccentricity from the fovea, and 2) the decrease in the accuracy of the eye tracker. The standard deviation of the Gaussian kernel used in our creation of saliency maps was equal to 2 degrees of visual angle.

V. Performance assessment

In order to assess the extent to which the depth saliency map can influence both the prediction of a saliency map and the overall performance of the proposed computational model, a set of quantitative comparisons between the fixation density map and the output of the proposed model are presented in this section.

A. Quantitative metrics of assessment

So far, there are no specific and standardized measures to compare the similarity between the fixation density maps and the saliency maps created by computational models in 3D situation. Nevertheless, there exists a range of different measures that are widely used to perform the comparison between saliency maps for 2D content. The most common ones include: (1) Pearson Linear Correlation Coefficient (PLCC) [53] [54], (2) Kullback-Leibler divergence (KLD) [53] [15], and (3) the area under the receiver operating characteristics curve (AUC) [18] [55]. The first two are directly applicable to a comparison between a fixation density map and a predicted saliency map, whereas AUC is usually applied to compare the actual fixation points to a predicted saliency map. Since the disparity compensation for binocular eye-tracking data has been done during the process of fixation density map creation,

TABLE II
PERFORMANCE OF THE DEPTH SALIENCY MAP (NOTED AS DSM IN THE TABLE), THE 2D SALIENCY MAPS FROM THREE 2D MODELS, THE DEPTH MAP AND THE DEPTH CONTRAST MAP. NOTE THAT A SMALLER KLD SCORE MEANS A BETTER PERFORMANCE. * MEANS THAT IT IS SIGNIFICANTLY DIFFERENT FROM THE PERFORMANCE OF THE DSM (PAIRED T-TEST, $p < 0.1$).

	PLCC	KLD	AUC
Itti's model	0.137*	2.819*	0.538*
Bruce's model	0.326	0.736	0.638
Hou's model	0.291	0.802*	0.630
Depth map	0.120*	1.022*	0.551*
Depth Contrast	0.181*	0.980*	0.586*
DSM	**0.368**	**0.708**	**0.656**

the two fixation density maps from both views have been merged into one. We therefore adopt these three similarity measures to quantitatively compare a merged fixation density map and a predicted saliency map obtained from one view.

B. Performance of depth saliency map

The creation of 2D saliency maps and of saliency maps based on depth information (i.e. depth saliency map) are the two main parts of the proposed model. In order to assess the extent to which these two sources of saliency maps can predict the salient areas in a scene, the performance of the depth saliency map (DSM) is compared with the performance of (1) the 2D saliency maps that comes from three state-of-the-art 2D visual attention models, and (2) the depth map and the depth contrast map which were used in existing models for predicting the 3D saliency. Here, the depth contrast map is obtained by the absolute values of the result of applying a DoG filter (the same one as used for computing DSM) on the depth map, based on the assumption that a depth contrast value of zero corresponds to zero saliency.

The results (see Table II) from all the three objective metrics show that the depth saliency map has a significantly higher performance than Itti's model, the depth map and the depth contrast map. Compared to Bruce's model and Hou's model, the performance of the depth saliency map is still higher, but without significant difference (except that the KL divergence value shows that the depth saliency map significantly outperforms Hou's model). These results demonstrate a great influence of the depth contrast on the distribution of visual attention in the viewing of 3D content. The performance of DSM, as compared to the depth contrast map, also illustrates the additional benefit of the proposed learning method.

C. Added value of a depth saliency map

The proposed model in the present paper belongs to the 'depth-saliency model' category, which highlights the existence of a depth saliency map. To compare the two different ways of making the most of depth information, the performance of the following methods were measured and compared:

- No-depth method. This is a direct use of a 2D computational model, no depth information is taken into account.
- Depth-weighting (DW) method. We adopt Chamaret's method [31], which weights each pixel in the 2D saliency map by multiplying it with the depth value of the corresponding pixel in the depth map (see Figure 1(a)). Since we do not have the code to apply exactly the same 2D computational model used in their paper, the 2D saliency map creation part is replaced by the models of Itti, Bruce, and Hou. This method is denoted as '2D $\times$ Depth' in Table III.
- Depth-saliency (DS) method, i.e. the proposed computational model in this paper. It creates a depth saliency map and a 2D saliency map respectively, then combines the resulting saliency maps from both paths (by equation 5). This method is denoted as '2D + DSM' in Table III.
- Other reference methods. For a fair comparison, we include also some other methods in the performance comparison, including (1) integrating the 2D saliency map with either a depth map or a depth contrast map using the proposed linear pooling strategy (denoted as '2D + Depth' and '2D + Depth Contrast', respectively); and (2) weighting the 2D saliency map by DSM (denoted as '2D * DSM').

TABLE III
CONTRIBUTION OF THE DEPTH INFORMATION ON 2D MODELS. + MEANS THE USE OF THE PROPOSED LINEAR POOLING STRATEGY INTRODUCED IN SECTION III-B. $\times$ MEANS THE WEIGHTING METHOD BASED ON MULTIPLICATION. NOTE THAT A SMALLER KLD SCORE MEANS A BETTER PERFORMANCE. * MEANS THAT IT IS SIGNIFICANTLY DIFFERENT FROM THE PERFORMANCE OF THE CORRESPONDING 2D MODEL (PAIRED T-TEST, $p < 0.1$).

		PLCC	KLD	AUC
Itti's model	2D model only	0.137	2.819	0.538
	2D + Depth	0.168	0.988*	0.567
	2D + Depth Contrast	0.211*	0.991*	0.596*
	2D $\times$ DSM	0.137	0.916	0.540
	2D $\times$ Depth (Chamaret)	0.137	2.916	0.540
	2D + DSM (Proposed)	**0.356***	**0.704***	**0.656***
Bruce's model	2D model only	0.326	0.736	0.638
	2D + Depth	0.282	0.792	0.621
	2D + Depth Contrast	0.343	0.669	0.644
	2D $\times$ DSM	0.403	0.686	0.671
	2D $\times$ Depth (Chamaret)	0.299	0.832	0.636
	2D + DSM (Proposed)	**0.424***	**0.617**	**0.675**
Hou's model	2D model only	0.291	0.802	0.630
	2D + Depth	0.246	0.848	0.607
	2D + Depth Contrast	0.307	0.711	0.362
	2D $\times$ DSM	0.341	0.782	0.660
	2D $\times$ Depth (Chamaret)	0.292	0.893	0.634
	2D + DSM (Proposed)	**0.410**	**0.605***	**0.670**
Upper Theoretical Similarity Limit		0.897	0.127	0.782

The performance of all these methods is shown in Table III. Additionally, in order to have an idea of what a good performance is, we compute the so-called Upper Theoretical Similarity Limit (UTPL) [56], which has been a common benchmark for 2D visual saliency models. The UTPL is computed as the similarity between the fixation density map obtained from half of the human observers (randomly selected) and the fixation density map resulting from the other half of the observers. We repeat this process 100 times to obtain a

robust estimate.

Large added values of the depth saliency map are demonstrated when it is combined with each of the three 2D visual attention models. The proposed model outperforms Chamaret's DW method and all the reference methods we introduced previously. Although there is still a considerable gap between the proposed model's performance and the UTPL, the proposed model has been demonstrated to have a level of performance on 3D content comparable to the performance of 2D models on 2D content To allow for a comparison, we remind here of the performance of the three state-of-the-art 2D models which has been validated on different 2D-image databases: Itti's model has a PLCC value ranging from 0.27 to 0.31 [57]; Bruce's model has a PLCC value ranging from 0.40 to 0.45 [57]; and Hou's model has an AUC value staying at around 0.69 [17].

D. Content-based analysis

The proposed database provides 3D images of different types of natural scenes. The variation in performance of the depth saliency map and its added value to 2D models makes a content-based analysis rather meaningful. For simplicity, (1) only Bruce's model is used as the reference to evaluate the performance and the added value of the depth saliency map; and (2) we adopt only the PLCC scores for this content-based analysis. Bruce's model is selected since it shows a relatively good performance on various types of scenes (a PLCC value ranging from 0.40 to 0.45 on different 2D image datasets [57]). The results are shown in Table IV. In most cases, the proposed method has significantly better results than either the 2D saliency map or the depth saliency map. However, the depth-weighting method (a multiplication of 2D saliency and depth map), obtains the best result for only one scene. In this scene (image 12 "Hall"), all the potentially salient areas have already been detected by the 2D visual attention model.

In order to further investigate the influence of DSM, an analysis regarding the performance and the added value of the DSM is performed. We compute the difference of the PLCC value for each image by Equation 7 and Equation 8:

$$\Delta^1_{PLCC} = PLCC_{DSM} - PLCC_{2D} \tag{7}$$

$$\Delta^2_{PLCC} = PLCC_{combined} - PLCC_{2D} \tag{8}$$

where $PLCC_{DSM}$ represents the performance of the depth saliency map (the second column in table IV), $PLCC_{2D}$ represents the performance of the 2D model (the first column in table IV), $PLCC_{combined}$ represents the performance of the proposed method (the fourth column in table IV). Therefore, Δ^1_{PLCC} indicates a relative performance of the depth saliency map compared to the 2D model, while Δ^2_{PLCC} indicates the added value. The results are plotted in figure 8.

In Figure 8, one can observe a linear relationship between the performance of the depth saliency map and its added value. A higher performance of the depth saliency map corresponds to a higher added value. Image clustering patterns can also be clearly observed: (1) four images are located in the region of higher performance of depth saliency map and high added value, (2) two images are placed in the region of lower performance and low (or even negative) added value, and (3) the remaining twelve images are spread around a comparable performance between depth saliency map and 2D model, with a considerable added value.

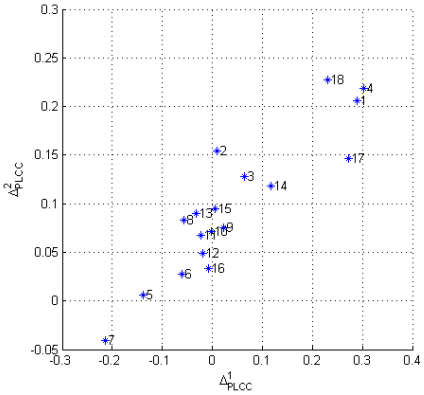


Fig. 8. The scatter plot of the performance and added value of the18 images. The ID of each image is indicated in Figure 5 and Figure 7.

When taking a closer look at the four images with high performance and high added value, one can observe that these images contain a huge amount of texture or salient 2D visual features. In Image 1 and Image 4, the widespread presence of faces and artificial color attracts the viewer's attention to most of the areas in the scene. On the other hand, in image 17 (Tree Branches), one can find a great amount of texture with similar color and no presence of any object of interest. We make a hypothesis that, when too many or too few areas are detected as salient by 2D visual features, it is difficult for 2D models to detect the 'real' salient region in a scene. Depth features might thus become an efficient cue to predict the viewer's attention. Image 18 represents a typical type of 3D scene: the main actor (or object) is given a positive disparity in order to create the "popping out" effect. In this situation, observers' attention is attracted by the popping out object. In this situation, the depth feature is obviously the dominant feature directing the observers' visual attention.

Image 5 and Image 7 are the two which yield the worst performance and lowest added value of DSM. In Image 5, the appearance of top-down features (text on the packing of the laundry detergent) attracts much attention. In Image 7, most of the objects of interest are with a color different from their background. These objects are located among some other objects of no-interest with the colors similar to the background. This special setup of a scene facilitates the 2D visual attention model to detect the salient areas. However, in terms of depth contrast, all the objects in the scene are considered as salient. The performance of the model is thus decreased by the increasing number of 'false-positive' detections. This problem regarding the increasing number of 'false-positive' implies that a normalized step (e.g. the one proposed by Itti *et al.* [14]) might be helpful.

For the remaining twelve images, the performances of DSM and 2D saliency maps are comparable. However, they also yield a considerable added value. The common characteristic of these images is that they generally have a small number of salient areas, which can be caused by either 2D salient

TABLE IV

PERFORMANCE (BASED ON THE METRIC PLCC) OF THE 2D MODEL, THE DEPTH SALIENCY MAP, AND THE ADDED VALUE OF DEPTH ACHIEVED BY (CHAMARET'S) DEPTH-WEIGHTING METHOD AND THE PROPOSED METHOD. * MEANS THAT IT IS SIGNIFICANTLY DIFFERENT FROM BRUCE'S 2D MODEL (PAIRED T-TEST, $p < 0.01$). FOR THE PROPOSED MODEL (I.E. COLUMN 5), $^{\#}$ MEANS THAT IT IS SIGNIFICANTLY DIFFERENT FROM THE DEPTH SALIENCY MAP (PAIRED T-TEST, $p < 0.01$). NOTE THAT THE ID OF EACH IMAGE IS INDICATED IN FIGURE 5 AND FIGURE 7.

ID	2D model (Bruce)	Depth saliency map	Chamaret's method	Proposed method
Image 1	0.113	**0.402***	0.042*	0.319*$^{\#}$
Image 2	0.364	0.373	0.512*	**0.519***$^{\#}$
Image 3	0.321	0.384*	0.231*	**0.449***$^{\#}$
Image 4	0.240	**0.542***	0.247	0.459*$^{\#}$
Image 5	0.252	0.114*	0.209*	**0.258**$^{\#}$
Image 6	0.568	0.507*	0.532*	**0.595***$^{\#}$
Image 7	**0.413**	0.198*	0.394*	0.372*$^{\#}$
Image 8	0.447	0.390*	0.376*	**0.531***$^{\#}$
Image 9	0.379	0.401*	0.336*	**0.454***$^{\#}$
Image 10	0.271	0.269	0.272	**0.343***$^{\#}$
Image 11	0.345	0.322*	0.159*	**0.413***$^{\#}$
Image 12	0.321	0.302*	**0.439***	0.370*$^{\#}$
Image 13	0.501	0.469*	0.272*	**0.591***$^{\#}$
Image 14	0.344	0.462*	0.291*	**0.462***
Image 15	0.513	0.517	0.509	**0.607***$^{\#}$
Image 16	0.232	0.225	0.232	**0.265***$^{\#}$
Image 17	-0.134	**0.139***	-0.062*	0.013*$^{\#}$
Image 18	0.367	0.598*	**0.603***	0.595*

features or depth. Therefore, the saliency map generated based on either 2D salient features or depth might predict parts of the salient area, but not all of them. This can be the reason why 2D saliency maps and depth saliency maps have comparable performances, but their combination has a much better result.

VI. DISCUSSION

A. Impact of the weighting

In the present study, we adopt a linear pooling strategy which equally weights the contributions of 2D and depth information to the final saliency map, since it is not yet fully understood how these two sources of saliency interact and finally affect the saliency distribution. In order to verify the extent to which the performance of the proposed model varies with weights w_1 and w_2 (i.e. various relative importance of depth and 2D information in saliency detection, respectively), we draw the curves of performance as a function of w_1 (see the Figure 9). We found that the maximal performance is achieved when w_1 approximates 0.6 (i.e. w_2 approximates 0.4), which means that depth information and 2D information may have comparable importance. This result supports our suggestion to consider depth information as an individual visual channel in modeling visual attention. Nevertheless, it is worth noting that this result is obtained based on only the proposed dataset. Moreover, according to different similarity metrics, the best performance is yielded by different w_1 values. Here, we would like to leave this issue of weighting as an open question and recommend $w_1 = w_2 = 0.5$ for the application of the proposed model.

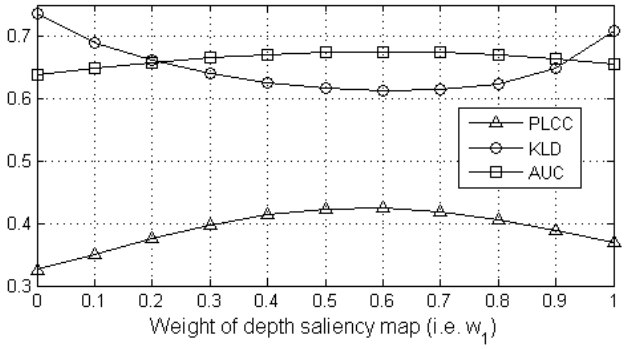


Fig. 9. Performance of different weighting.

B. Issues related to eye-tracking experiment

In our eye-tracking experiment, the apparatus used can only provide a two-dimensional spatial gaze location for each eye separately. We found out that disparity exists between the left fixation and the right fixation. This disparity is related to the local image disparity, which means that the participants indeed fixated in depth. However, the triangulation of the two 2D gaze points from both eyes to produce a single 3D gaze point is not straightforward, and has not been fully understood. Moreover, the use of this triangulation also relies on the calibration of the system. For an experiment using 2D stimuli, calibration points are typically shown on the screen. It is thus easy to determine whether the observer is looking accurately at the point or not, since the 2D coordinates are known and the 2D gaze can be accurately tracked on the screen plane. However, an experiment using 3D stimuli requires a volumetric calibration (e.g. by showing points at different depth planes) in order to compute the 3D gaze points [23].

In addition to the calibration, the presentation time of each stimulus is another crucial factor in conducting an eye-tracking experiment. So far, there are no standardized methodologies for the conduction of eye-tracking experiments for 3D images.

In our experiment, the presentation time was set to 15 seconds, in order to avoid losing useful data. However, the 15-second presentation time is relatively long as compared to the ones used in 2D eye-tracking. To analyze the impact of the presentation time on the resulting ground truth fixation density map, we evaluate the performance of computational models using PLCC based on the fixation density maps obtained with different viewing durations (ranging from 1 to all the 15 seconds). The results are illustrated in Figure 10.

Generally, it is believed that the influence of bottom-up mechanisms is strong for early fixations. We also found out a strong impact of center-bias for a short observation duration. This might be due to the presence of a central point between two stimuli. Therefore, comparing the output of the models with these ground truths might not illustrate the model's real performance (see Figure 10).

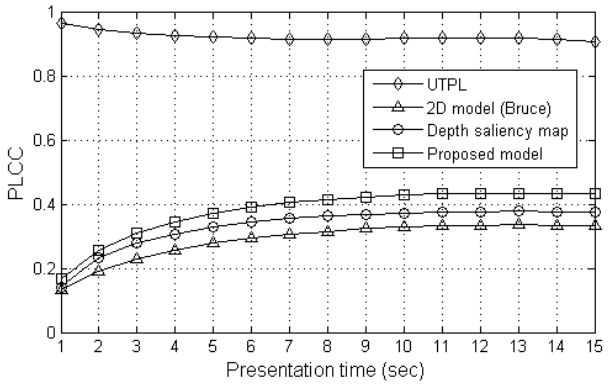


Fig. 10. Performance of models as a function of presentation time.

As the observation time increases, the impact of top-down mechanisms increases. Nevertheless, it is believed that the impact of bottom-up mechanisms does not disappear; both types of mechanisms interact and compete with each other to have an impact on visual attention [3]. As seen in Figure 10, the evaluated performance of the models does not decrease with a longer presentation time, despite that the models are believed to be designed based on bottom-up cues. Using the ANOVA test, we found that the performance of each of the models does not show any significant differences after 10 seconds (for 2D saliency map, $p = 0.798$; for the DSM, $p = 0.775$; for the combination, $p = 0.637$). For the proposed eye-tracking database, we therefore argue that using a presentation time less than 10 seconds to create the ground truth fixation density maps might affect the evaluation of the model's performance, but using a presentation time longer than 10 seconds has no negative impact.

VII. Conclusion and perspective

In this paper, we have presented a depth-feature based computational model of visual attention for 3D still images. The proposed model contains a depth saliency map creation part which is based on a probability-learning from eye-tracking data. We have also created a database which contains eighteen stereoscopic 3D images with natural content, disparity maps, and free-task viewing eye-tracking data. The creation of this database enables the quantitative evaluation of the proposed model and also solves the problem of the lack of ground truth in the area of 3D visual attention modeling.

Our study has shown depth contrast as a saliency cue that directs the observer's attention in the viewing of 3D still images. By merging the depth saliency map with the results of 2D visual feature detection, the proposed model yields a good prediction of salient areas. The performance of the proposed model on 3D image is comparable to the performance of state-of-the-art 2D models on 2D images. Moreover, we have shown that the performance of the depth saliency map and its added value to a 2D model vary across different types of scene.

We have also compared various ways of applying depth information to 3D visual attention models. Our result indicates the importance of a depth saliency map in the modeling of 3D visual attention. Nevertheless, this result should not lead to a strong conclusion that a depth-saliency model is definitely better or worse than a depth-weighting model, since the depth-weighting model has various advantages too, such as low computation complexity or comparable performance for some types of scenes. On the other hand, it would be reasonable to suggest that an efficient 3D visual attention model can be a combination of both types of models: firstly, depth information is processed as an additional visual dimension from which depth features are extracted to create depth saliency maps; secondly, depth can be also used as weighting information to relate the distribution of attention to the distance between observer and each object in the scene [27].

In the present study, even if its performance is good, our model still suffers from some limitations. The main one is that the proposed model only exploits depth contrast. A potential way to improve the proposed model relies on additional depth features. In the literature, several depth features have been proposed and investigated, such as surface curvature, depth gradient and orientation contrast [32] [58] [59]. However, since it has been demonstrated that the influence of these features might largely differ from one another [41], the application of more depth features raises the demand for a potential normalization step for each feature dimension and a more sophisticated pooling strategy. In our future work, we will evaluate the effects of different depth features and try to extend the proposed model by taking into account more depth features.

(This work is supported by the French ANR-PERSEE project ANR-09-BLAN-0170.)

References

[1] J. Wolfe, "Visual attention," *Seeing*, vol. 2, pp. 335–386, 2000.
[2] A. Yarbus, *Eye movements and vision*. Plenum press, 1967.
[3] J. Wang, D. M. Chandler, and P. L. Callet, "Quantifying the relationship between visual salience and visual importance," ser. SPIE Proceedings, vol. 7527. SPIE, 2010, p. 75270.
[4] W. Einhäuser, U. Rutishauser, and C. Koch, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *Journal of Vision*, vol. 8, no. 2, 2008.
[5] J. Henderson, J. Brockmole, M. Castelhano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes," *Eye movements: A window on mind and brain*, pp. 537–562, 2007.
[6] W. van Zoest and M. Donk, "Bottom-up and top-down control in visual search," *PERCEPTION-LONDON-*, vol. 33, pp. 927–938, 2004.

[7] J. Theeuwes, "Exogenous and endogenous control of attention: The effect of visual onsets and offsets," *Attention, Perception, & Psychophysics*, vol. 49, no. 1, pp. 83–90, 1991.

[8] J. Wolfe, S. Butcher, C. Lee, and M. Hyle, "Changing your mind: on the contributions of top-down and bottom-up guidance in visual search for feature singletons." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, p. 483, 2003.

[9] K. Park and H. Park, "Region-of-interest coding based on set partitioning in hierarchical trees," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 2, pp. 106–113, 2002.

[10] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, "Retargeting images and video for preserving information saliency," *IEEE Computer Graphics and Applications*, pp. 80–88, 2007.

[11] K. Vu, K. Hua, and W. Tavanapong, "Image retrieval based on regions of interest," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 1045–1049, 2003.

[12] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 7, pp. 971–982, 2011.

[13] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[14] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[15] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

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

[17] O. Le Meur and J. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2801–2813, 2010.

[18] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008.

[19] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Advances in neural information processing systems*, vol. 18, p. 547, 2006.

[20] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[21] J. Hakkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman, "What do people look at when they watch stereoscopic movies?" J. W. Andrew, S. H. Nicolas, and A. D. Neil, Eds., vol. 7524. SPIE, 2010, p. 75240E.

[22] Q. Huynh-Thu and L. Schiatti, "Examination of 3d visual attention in stereoscopic video content," in *Proceedings of SPIE*, vol. 7865, 2011, p. 78650J.

[23] 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*, vol. 57, no. 2, pp. 421–431, 2011.

[24] L. Jansen, S. Onat, and P. König, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *Journal of Vision*, vol. 9, no. 1, 2009.

[25] Y. Liu, L. Cormack, and A. Bovik, "Natural scene statistics at stereo fixations," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 161–164.

[26] C. Ramasamy, D. House, A. Duchowski, and B. Daugherty, "Using eye tracking to analyze stereoscopic filmmaking," in *SIGGRAPH'09: Posters*. ACM, 2009, p. 28.

[27] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel, and M. Perreira Da Silva, "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli," *Journal of Eye Movement Research*, vol. 5(5):1, pp. 1–11, 2012.

[28] D. Wismeijer, C. Erkelens, R. van Ee, and M. Wexler, "Depth cue combination in spontaneous eye movements," *Journal of vision*, vol. 10, no. 6, 2010.

[29] A. Maki, P. Nordlund, and J. Eklundh, "A computational model of depth-based attention," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 4. IEEE, 1996, pp. 734–739.

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

[31] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, "Adaptive 3d rendering based on region-of-interest," in *Proceedings of SPIE*, vol. 7524, 2010, p. 75240V.

[32] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 1. IEEE, 2000, pp. 375–378.

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

[34] N. Bruce and J. Tsotsos, "An attentional framework for stereo vision," in *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*. IEEE, 2005, pp. 88–95.

[35] D. Hoffman, A. Girshick, K. Akeley, and M. Banks, "Vergence–accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3, 2008.

[36] Y. Okada, K. Ukai, J. Wolffsohn, B. Gilmartin, A. Iijima, and T. Bando, "Target spatial frequency determines the response to conflicting defocus-and convergence-driven accommodative stimuli," *Vision Research*, vol. 46, no. 4, pp. 475–484, 2006.

[37] J. Tsotsos, S. Culhane, W. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[38] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A bayesian inference theory of attention," *Vision research*, vol. 50, no. 22, pp. 2233–2247, 2010.

[39] S. Pinneli and D. Chandler, "A bayesian approach to predicting the perceived interest of objects," in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 2584–2587.

[40] A. Brookes and K. Stevens, "The analogy between stereo depth and brightness," *Perception*, vol. 18, no. 5, pp. 601–614, 1989.

[41] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H. Seidel, "A perceptual model for disparity," *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011, Vancouver)*, vol. 30, no. 4, 2011.

[42] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New requirements of subjective video quality assessment methodologies for 3dtv," in *Video Processing and Quality Metrics 2010 (VPQM), Scottsdale, USA, 2010.*, 2010.

[43] J. Wang, M. Barkowsky, V. Ricordel, P. Le Callet *et al.*, "Quantifying how the combination of blur and disparity affects the perceived depth," in *Proc. SPIE 7865, 78650K (2011)*, 2011.

[44] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[45] 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.

[46] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[47] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–195.

[48] M. Halle, "Autostereoscopic displays and computer graphics," in *ACM SIGGRAPH 2005 Courses*. ACM, 2005, p. 104.

[49] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–721.

[50] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, and N. Garcia, "Nama3ds1-cospad1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, july 2012, pp. 109 –114.

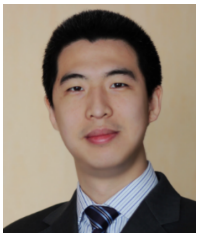
[51] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2464–2471.

[52] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic huber-l1 optical flow," in *Proceedings of the British machine vision conference*, 2009.

[53] 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*, vol. 28, no. 5, pp. 802–817, 2006.

[54] U. Engelke, A. Maeder, and H. Zepernick, "Analysing inter-observer saliency variations in task-free viewing of natural images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1085–1088.

[55] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, 2011.
[56] B. Stankiewicz, N. Anderson, and R. Moore, "Using performance efficiency for testing and optimization of visual attention models," in *Proceedings of SPIE*, vol. 7867, 2011, p. 78670Y.
[57] M. Perreira Da Silva, "Modèle computationnel d'attention pour la vision adaptative," Ph.D. dissertation, 2010.
[58] G. Kootstra, N. Bergstrom, and D. Kragic, "Fast and automatic detection and segmentation of unknown objects," in *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*. IEEE, 2010, pp. 442–447.
[59] A. Jain and C. Kemp, "El-e: an assistive mobile manipulator that autonomously fetches objects from flat surfaces," *Autonomous Robots*, vol. 28, no. 1, pp. 45–64, 2010.



Junle Wang received the double-M.S. degree in Signal processing (South China University of Technology, China), and Electronic Engineering (University of Nantes, France) in 2009, and the PhD degree in computer science from University of Nantes in 2012. He is currently an ATER (Assistant Professor) at the Department of Electronic and Digital Technologies, Ecole polytechnique de l'universidé de Nantes. His research focuses on visual attention, depth perception and quality of experience of stereoscopic-3D content. His research interests also include quality assessment, human visual perception and psychophysical experimentation.

Matthieu Perreira Da Silva is associate professor in the IVC team of the IRRCyN Lab since September 2011. Apart from his research activities in the team, he is also teaching in the computer science department of Polytech Nantes engineering school. He received a M.Sc in image processing in 2001 and a Ph.D. in computer science and applications in 2010, both from the University of La Rochelle, France. From 2001 to 2006 he worked as a R&D engineer in Zefyr Technologies, a private company dealing with biometric identification. From 2006 to mid 2011, he was successively engineer, Ph.D. student and teaching assistant at the University of La Rochelle. His research interests include human perception, visual attention, human computer interaction, artificial curiosity, autonomous machine learning, image processing and computer vision.

Patrick Le Callet received M.Sc. degree PhD degree in image processing from Ecole polytechnique de l'Université de Nantes. He was also student at the Ecole Normale Superieure de Cachan where he get the "Aggrégation" (credentialing exam) in electronics of the French National Education. He has been working as an Assistant professor from 1997 to 1999 and as a full time lecturer from 1999 to 2003 at the department of Electrical engineering of Technical Institute of University of Nantes (IUT). Since 2003 he is teaching at Ecole polytechnique de l'Universié de Nantes (Engineer School) in the Electrical Engineering and the Computer Science department where is now Full Professor. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. His current centers of interest are 3D image and video quality of experience, watermarking techniques and visual attention modeling and applications. He is co-author of more than 140 publications and communications and co-inventor of 13 international patents on these topics.

Vincent Ricordel received the Ph.D degree in Signal Processing and Telecommunication from the University of Rennes, Frances, in 1996. After a post-doct in 1997 at the Tampere University of Technology, Finland, he has been, from 1998 to 2002, an associate professor at the University of Toulon, France. Since 2002, he holds an associate professor position at the Polytechnic school of the University of Nantes, France, where he is now the head of the computer science department. He is also a member of the IVC team of the IRCCyN laboratory. His research interests include video coding, image sequence analysis, vector quantization, and visual attention.