Saliency Detection for Stereoscopic Images

Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, Weisi Lin

Saliency Detection for Stereoscopic Images

Yuming Fang, *Member, IEEE*, Junle Wang, Manish Narwaria, Patrick Le Callet, *Member, IEEE*, and Weisi Lin, *Senior Member, IEEE*

***Abstract*— Many saliency detection models for 2D images have been proposed for various multimedia processing applications during the past decades. Currently, the emerging applications of stereoscopic display require new saliency detection models for salient region extraction. Different from saliency detection for 2D images, the depth feature has to be taken into account in saliency detection for stereoscopic images. In this paper, we propose a novel stereoscopic saliency detection framework based on the feature contrast of color, luminance, texture, and depth. Four types of features, namely color, luminance, texture, and depth, are extracted from discrete cosine transform coefficients for feature contrast calculation. A Gaussian model of the spatial distance between image patches is adopted for consideration of local and global contrast calculation. Then, a new fusion method is designed to combine the feature maps to obtain the final saliency map for stereoscopic images. In addition, we adopt the center bias factor and human visual acuity, the important characteristics of the human visual system, to enhance the final saliency map for stereoscopic images. Experimental results on eye tracking databases show the superior performance of the proposed model over other existing methods.**

***Index Terms*— Stereoscopic image, 3D image, stereoscopic saliency detection, visual attention, human visual acuity.**

I. 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 part by filtering out others to reduce the complexity of 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 in 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 [1]–[8], while top-down approach is a task-dependent cognitive processing affected by the performed tasks, feature distribution of targets, etc. [9]–[11].

Manuscript received June 17, 2013; revised November 14, 2013 and January 7, 2014; accepted January 26, 2014. Date of publication February 6, 2014; date of current version May 9, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Damon M. Chandler.

Y. Fang is with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China (e-mail: fa0001ng@e.ntu.edu.sg).

J. Wang, M. Narwaria, and P. Le Callet are with LUNAM Université, Université de Nantes, Nantes Cedex 3 44306, France (e-mail: wang.junle@gmail.com; mani0018@e.ntu.edu.sg; patrick.lecallet@univ-nantes.fr).

W. Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2305100

Over the past decades, many studies have tried to propose computational models of visual attention for various multimedia processing applications, such as visual retargeting [5], visual quality assessment [9], [13], visual coding [14], etc. In these applications, the salient regions extracted from saliency detection models are processed specifically since they attract much more humans' attention compared with other regions. Currently, many bottom-up saliency detection models have been proposed for 2D images/videos [1]–[8].

Today, with the development of stereoscopic display, there are various emerging applications for 3D multimedia such as 3D video coding [31], 3D visual quality assessment [32], [33], 3D rendering [20], etc. In the study [33], the authors introduced the conflict met by the HVS while watching 3D-TV, how these conflicts might be limited and how visual comfort might be improved by the visual attention model. The study also described some other visual attention based 3D multimedia applications, which exist in different stages of a typical 3D-TV delivery chain, such as 3D video capture, 2D to 3D conversion, reframing and depth adaptation, etc. Chamaret *et al.* adopted ROIs for 3D rendering in the study [20]. Overall, the emerging demand of visual attention based applications for 3D multimedia increases the requirement of computational saliency detection models for 3D multimedia content.

Compared with various saliency detection models proposed for 2D images, only a few studies exploiting the 3D saliency detection exist currently [18]–[27]. Different from saliency detection for 2D images, the depth factor has to to be considered in saliency detection for 3D images. To achieve the depth perception, binocular depth cues (such as binocular disparity) are introduced and merged together with others (such as monocular disparity) in an adaptive way based on the viewing space conditions. However, this change of depth perception also largely influences the human viewing behavior [39]. Therefore, how to estimate the saliency from depth cues and how to combine the saliency from depth with those from other 2D low-level features are two important factors in designing 3D saliency detection models.

In this paper, we propose a novel saliency detection model for 3D images based on feature contrast from color, luminance, texture, and depth. The features of color, luminance, texture and depth are extracted from DCT (Discrete Cosine Transform) coefficients of image patches. It is well accepted that the DCT is a superior representation for energy compaction and most of the signal information is concentrated on a few low-frequency components [34]. Due to its energy compactness property, the DCT has been widely used in various signal

processing applications in the past decades. Our previous study has also demonstrated that DCT coefficients can be adopted for effective feature representation in saliency detection [5]. Therefore, we use DCT coefficients for feature extraction for image patches in this study.

In essence, the input stereoscopic image and depth map are firstly divided into small image patches. Color, luminance and texture features are extracted based on DCT coefficients of each image patch from the original image, while depth feature is extracted based on DCT coefficients of each image patch in the depth map. Feature contrast is calculated based on center-surround feature difference, weighted by a Gaussian model of spatial distances between image patches for the consideration of local and global contrast. A new fusion method is designed to combine the feature maps to obtain the final saliency map for 3D images. Additionally, inspired by the viewing influence from centre bias and the property of human visual acuity in the HVS, we propose to incorporate the centre bias factor and human visual acuity into the proposed model to enhance the saliency map. The Centre-Bias Map (CBM) calculated based on centre bias factor and a statistical model of human visual sensitivity in [38] are adopted to enhance the saliency map for obtaining the final saliency map of 3D images. Existing 3D saliency detection models usually adopt depth information to weight the traditional 2D saliency map [19], [20], or combine the depth saliency map and the traditional 2D saliency map simply [21], [23] to obtain the saliency map of 3D images. Different from these existing methods, the proposed model adopts the low-level features of color, luminance, texture and depth for saliency calculation in a whole framework and designs a novel fusion method to obtain the saliency map from feature maps. Experimental results on eye-tracking databases demonstrate the superior performance of the proposed model over other existing methods.

The remaining of this paper is organized as follows. Section II introduces the related work in the literature. In Section III, the proposed model is described in detail. Section IV provides the experimental results on eye tracking databases. The final section concludes the paper.

II. Related Work

As introduced in the previous section, many computational models of visual attention have been proposed for various 2D multimedia processing applications. Itti *et al.* proposed one of the earliest computational saliency detection models based on the neuronal architecture of the primates' early visual system [1]. In that study, the saliency map is calculated by feature contrast from color, intensity and orientation. Later, Harel *et al.* extended Itti's model by using a more accurate measure of dissimilarity [2]. In that study, the graph-based theory is used to measure saliency from feature contrast. Bruce *et al.* designed a saliency detection algorithm based on information maximization [3]. The basic theory for saliency detection is Shannon's self-information measure [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 [12].

Hou *et al.* proposed a saliency detection method by the concept of Spectral Residual [4]. The saliency map is computed by log spectra representation of images 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 [14]. Yan *et al.* introduced a saliency detection algorithm based on sparse coding [8]. Recently, some saliency detection models have been proposed by patch-based contrast and obtain promising performance for salient region extraction [5]–[7]. Goferman *et al.* introduced a context-aware saliency detection model based on feature contrast from color and intensity in image patches [7]. A saliency detection model in compressed domain is designed by Fang *et al.* for the application of image retargeting [5].

Besides 2D saliency detection models, several studies have explored the saliency detection for 3D multimedia content. In [18], Bruce *at al.* proposed a stereo attention framework by extending an existing attention architecture to the binocular domain. However, there is no computational model proposed in that study [18]. Zhang *et al.* designed a stereoscopic visual attention algorithm for 3D video based on multiple perceptual stimuli [19]. Chamaret *et al.* built a Region of Interest (ROI) extraction method for adaptive 3D rendering [20]. Both studies [19] and [20] 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 [21], Ouerhani *et al.* extended a 2D saliency detection model to 3D saliency detection by taking depth cues into account. Potapova introduced a 3D saliency detection model for robotics tasks by incorporating the top-down cues into the bottom-up saliency detection [22]. Lang *et al.* conducted eye tracking experiments over 2D and 3D images for depth saliency analysis and proposed 3D saliency detection models by extending previous 2D saliency detection models [26]. Niu *et al.* explored the saliency analysis for stereoscopic images by extending a 2D image saliency detection model [25]. Ciptadi *et al.* used the features of color and depth to design a 3D saliency detection model for the application of image segmentation [27]. Recently, Wang *et al.* proposed a computational model of visual attention for 3D images by extending the traditional 2D saliency detection methods. In the study [23], the authors provided a public database with ground-truth of eye-tracking data.

From the above description, the key of 3D saliency detection model is how to adopt the depth cues 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. [15]–[17]. Therefore, an accurate 3D saliency detection model should take depth contrast into account as well as contrast from other common 2D low-level features. Accordingly, we propose a saliency

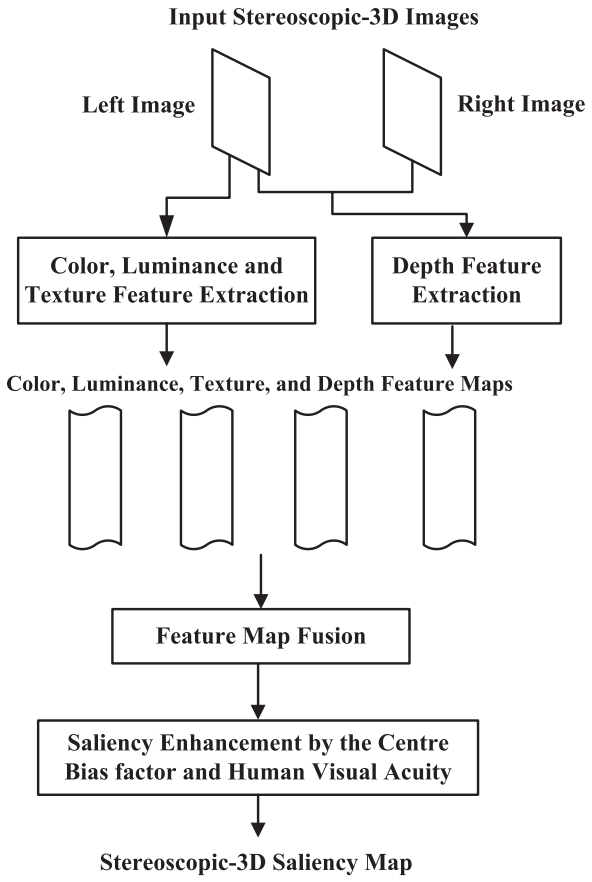


Fig. 1. The framework of the proposed model.

detection framework based on the feature contrast from low-level features of color, luminance, texture and depth. A new fusion method is designed to combine the feature maps for the saliency estimation. Furthermore, the centre bias factor and the human visual acuity are adopted to enhance the saliency map for 3D images. The proposed 3D saliency detection model can obtain promising performance for saliency estimation for 3D images, as shown in the experiment section.

III. The Proposed Model

The framework of the proposed model is depicted as Fig. 1. Firstly, the color, luminance, texture, and depth features are extracted from the input stereoscopic image. Based on these features, the feature contrast is calculated for the feature map calculation. A fusion method is designed to combine the feature maps into the saliency map. Additionally, we use the centre bias factor and a model of human visual acuity to enhance the saliency map based on the characteristics of the HVS. We will describe each step in detail in the following subsections.

A. 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. Our experimental results show that the proposed model with the patch size within the visual angle of [0.14, 0.21] (degrees) can get promising performance. In this paper, we use the patch size of 8×8 (the visual angle within the range of [0.14, 0.21] degrees) for the saliency calculation. The used image patch size is also the same as DCT block size in JPEG compressed images. The input RGB image is 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-opponent components. For the DCT coefficients, DC coefficients represent the average energy over all pixels in the image patch, while AC coefficients represent the detailed frequency properties of 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 features 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 and little texture information is included in these two channels, we use AC coefficients from only Y component to represent the texture feature of the image patch. In DCT block, most of the energy is included in the first several low-frequency coefficients in the left-upper corner of the DCT block. As there is little energy with the high-frequency coefficients in the right-bottom corner of the DCT block, we just use several first AC coefficients to represent the texture feature of image patches. The existing study in [35] demonstrates 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 [35], we use the first 9 low-frequency AC coefficients to represent the texture feature for each image patch as $T = \{Y_{AC1}, Y_{AC2}, \ldots, Y_{AC9}\}$.

For the depth feature, we assume that a depth map provides the information of the perceived depth for the scene. In a stereoscopic display system, depth information is usually represented by a disparity map which shows the parallax of each pixel between the left-view and the right-view images. The disparity is usually measured in unit of pixels for display systems. In this study, the depth map M of perceived depth information is computed based on the disparity as [23]:

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

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. We set the parameters based on the experimental studies in [23].

Similar with feature extraction for color and luminance, we adopt the DC coefficients of patches in depth map calculated in Eq. (1) as $D = M_{DC}$ (M_{DC} represents the DC coefficient of 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.

B. Feature Map Calculation

As we have explained before, salient regions in visual scenes pop out due to their feature contrast from their surrounding regions. Thus, a direct method to extract salient regions in visual scenes is to calculate the feature contrast between image patches and their surrounding patches in visual scenes. In this study, we estimate the saliency value of each image patch based on the feature contrast between this image path and all the other patches in the image. Here, we use a Gaussian model of spatial distance between image patches to weight the feature contrast for saliency calculation. The saliency value F_i^k of image patch i from feature k can be calculated as:

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

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 of the Gaussian model and it determines the degree of local and global contrast for the saliency estimation. σ is set as 5 based on the experiments of the previous work [5]. For any image patch i, its saliency value is calculated based on the center-surround differences between this patch and all other patches in the image. The weighting for the center-surround differences is determined by the spatial distances (within the Gaussian model) between image patches. The differences from nearer image patches will contribute more to the saliency value of patch i than those from farther image patches. Thus, we consider both local and global contrast from different features in the proposed saliency detection model.

The feature difference U_{ij}^k between image patches i and j is computed differently from features k due to the different feature representation method. Since the color, luminance and depth features are represented by one DC coefficient for each image patch, the feature contrast from these features (luminance, color and depth) between two image patches i and j can be calculated as the difference between two DC coefficients of two corresponding image patches as follows.

$$U_{ij}^m = \frac{|B_i^m - B_j^m|}{B_i^m + B_j^m} \tag{3}$$

where B^m represents the feature and $B^m \in \{L, C_1, C_2, D\}$; the denominator is used to normalize the feature contrast.

Since texture feature is represented as 9 low-frequency AC coefficients, we calculate the feature contrast from texture by the L2 norm. The feature contrast $U_{ij}^{'}$ from texture feature between two image patches i and j can be computed as follows.

$$U_{ij}^{'} = \frac{\sqrt{\sum_t (B_i^{'t} - B_j^{'t})^2}}{\sum_t (B_i^{'t} + B_j^{'t})} \tag{4}$$

where t represents the AC coefficients and $t \in \{1, 2, ..., 9\}$; $B^{'}$ represents the texture feature; the denominator is adopted to normalize the feature contrast.

C. Saliency Estimation from Feature Map Fusion

After calculating feature maps indicated in Eq. (2), we fuse these feature maps from color, luminance, texture and depth to compute the final saliency map. It is well accepted that different visual dimensions in natural scenes are competing with each other during the combination for the final saliency map [40], [41]. Existing studies have shown that a stimulus from several saliency features is generally more conspicuous than that from only one single feature [1], [41]. The different visual features interact and contribute simultaneously to the saliency of visual scenes. Currently, existing studies of 3D saliency detection (*e.g.* [23]) use simple linear combination to fuse the feature maps to obtain the final saliency map. The weighting of the linear combination is set as constant values and is the same for all images. To address the drawbacks from ad-hoc weighting of linear combination for different feature maps, we propose a new fusion method to assign adaptive weighting for the fusion of feature maps in this study.

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. Thus, a good feature map should detect small and compact 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.

$$\upsilon_k = \frac{\sum_{(i,j)} \sqrt{(i - E_{i,k})^2 + (j - E_{j,k})^2} \cdot F_k(i, j)}{\sum_{(i,j)} F_k(i, j)} \tag{5}$$

where (i, j) is the spatial location in the feature map; k represents the feature channel and $k \in \{L, C_1, C_2, T, D\}$; $(E_{i,k}, E_{j,k})$ is the average spatial location weighted by feature response, which is calculated as:

$$E_{i,k} = \frac{\sum_{(i,j)} i \cdot F_k(i, j)}{\sum_{(i,j)} F_k(i, j)} \tag{6}$$

$$E_{j,k} = \frac{\sum_{(i,j)} j \cdot F_k(i, j)}{\sum_{(i,j)} F_k(i, j)} \tag{7}$$

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.

$$\beta_k = 1/(e^{\upsilon_k}) \tag{8}$$

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

Based on compactness property of feature maps calculated in Eq. (8), we fuse the feature maps for the saliency map as follows.

$$S_f = \sum_k \beta_k \cdot F_k + \sum_{p \neq q} \beta_p \cdot \beta_q \cdot F_p \cdot F_q \tag{9}$$

The first term in Eq. (9) represents the linear combination of feature maps weighted by corresponding compactness properties of feature maps; while the second term is adopted to

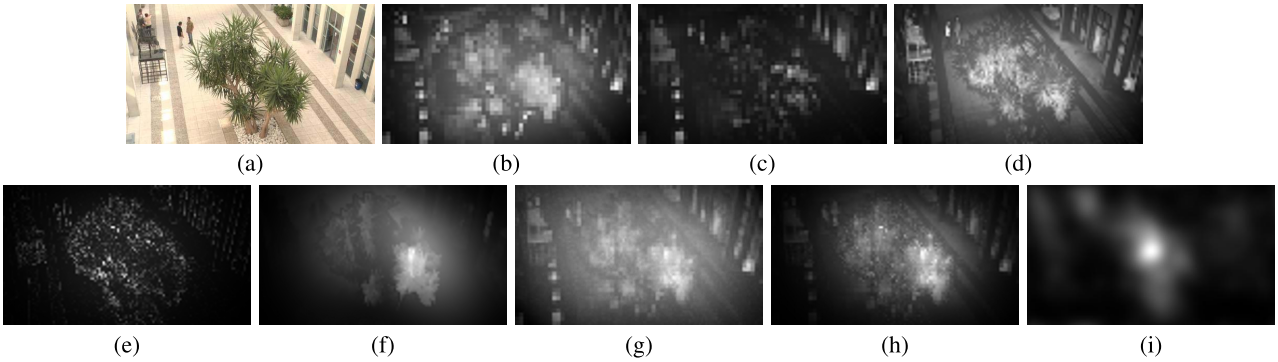


Fig. 2. Visual samples for different feature maps and saliency maps: (a) original image; (b) color feature map from C_b component; (c) color feature map from C_r component; (d) luminance feature map; (e) texture feature map; (f) depth feature map; (g) saliency map from linear combination of the feature maps with the same weighting; (h) saliency map from the proposed combination method (the weights of C_b color, C_r color, luminance, texture, and depth feature maps are 0.45, 0.51, 0.49, 0.62, and 0.81, respectively); (i) ground truth map.

enhance the common salient regions which can be detected by any two different feature maps. Different 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 properties. Fig. 2 provides an image sample for the feature map fusion. In this figure, Fig. 2(g) shows the saliency map by combing the feature maps with the same weighting; Fig. 2(h) gives the saliency map from the proposed combination method, which combine the feature maps with different weights. From this figure, we can see that the proposed combination method gives more weighting to the depth feature map during the fusion process of feature maps, which causes the final saliency map more similar with the ground truth map. Experimental results in the next section show that the proposed fusion method can obtain promising performance.

D. Saliency Enhancement

Eye tracking experiments from existing studies have shown that the bias towards the screen center exists during human fixation, which is called centre bias [43], [44]. In the study [43], the experiments show that the initial response is to orient to the screen center when the scene appears. The study [44] also shows that the center-bias exists during the human fixation. Existing studies have demonstrated that the performance of fixation prediction can be improved largely by considering the centre bias factor in saliency detection models [45], [46]. In this paper, we have used the centre bias factor to enhance the saliency map from the proposed 3D saliency detection model. Similarly with the studies [43], [45], [46], we use a Gaussian function with kernel width as one degree (foveal size) to model the centre bias factor. A CBM S_c can be obtained by the Gaussian function.

The experimental results in the study [43] shows the centre bias is irrespective to the distribution of image features, which means that the centre bias is independent on the saliency map S_f calculated from image features. Here, we consider the CBM as the fixation estimation from centre bias factor, similarly with the fixation estimation map S_f (saliency map) from image features. The saliency map by considering the center bias factor can be calculated as follows.

$$S = \gamma_1 S_f + \gamma_2 S_c \tag{10}$$

where γ_1 and γ_2 are two parameters used to weight the two components. In the experiment, we consider the saliency map S_f from image features more important than the CBM S_c from center-bias factor, and the parameters are set as $\gamma_1 = 0.7$ and $\gamma_2 = 0.3$ empirically.

It is well accepted that the HVS is highly space-variant due to the different densities of cone photoreceptor cells in the retina [36]. On the retina, the fovea owns the highest density of cone photoreceptor cells. Thus, the focused region has to be projected on the fovea to be perceived at the highest resolution. The density of the cone photoreceptor cells becomes lower with larger retinal eccentricity. The visual acuity decreases with the increased eccentricity from the fixation point [36], [38]. We use this property to enhance the saliency map of 3D images. In the saliency map, the pixels whose saliency value is larger than certain threshold are considered as salient regions. The human eyes would focus on these salient regions when observing the natural scenes and they are also most sensitive to these regions. The human visual acuity decreases with farther neighboring regions of these salient regions. In this study, we use a model of human visual sensitivity in [38] to weight the saliency map. The contrast sensitivity $C_s(f, e)$ can be calculated as [38]:

$$C_s(f, e) = \frac{1}{C_0 exp(\alpha f (e + e_2)/e_2)} \tag{11}$$

where f is the spatial frequency (cycles/degree); e is the retinal eccentricity (degree); C_0 is the minimum contrast threshold; α is the spatial frequency decay constant; e_2 is the half-resolution eccentricity. Based on the experimental results in [38], the best fitting parameter values are: $\alpha = 0.106$, $e_2 = 2.3$, $C_0 = 1/64$.

The retina eccentricity e between the salient pixel and non-salient pixel can be computed according to its relationship with spatial distance between image pixels. For any pixel position (i, j), its eccentricity e can be calculated by the spatial distance between this pixel and the nearest salient pixel (i_0, j_0) as:

$$e = tan^{-1}(d'/v) \tag{12}$$

(a) (b) (c) (d)

Fig. 3. Visual comparison samples between the original saliency map and enhanced saliency map by centre-bias and human visual acuity. (a) Input image. (b) Original saliency map. (c) Enhanced saliency map. (d) Ground truth map.

where v is the viewing distance; d' is the spatial distance between image pixels (i_0, j_0) and (i, j).

The final saliency map S' enhanced by the normalized visual sensitivity $C_s(f, e)$ can be calculated as:

$$S' = S * C_s(f, e) \tag{13}$$

With the enhancement operation by the centre bias factor, the saliency values of center regions in images would increase, while with the enhancement operation by human visual acuity, the saliency values of non-salient regions in natural scenes would decrease and the saliency map would get visually better. Fig. 3 provides one visual comparison sample between the original saliency map and the enhanced saliency map by the centre bias factor and human visual acuity. From this figure, we can see that the central regions become more salient with the enhancement by the centre bias factor. Additionally, the saliency values of non-salient regions in the saliency maps decreased by the enhancement operation of the human visual acuity. With the enhancement operation by the centre bias factor and human visual acuity, the saliency map can predict the saliency more accurately, as shown in Fig. 3, in which the enhanced saliency map (Fig. 3(c)) is more similar with the ground truth map (Fig. 3(d)) compared with the original saliency map (Fig. 3(b)). Please note that the ground truth map is obtained by the fixation data recorded from eye tracker [23].

IV. Experiment Evaluation

In this section, we conduct the experiments to demonstrate the performance of the proposed 3D saliency detection model. We first present the evaluation methodology and quantitative evaluation metrics. Following this, the performance comparison between different feature maps is given in subsection IV-B. In Subsection IV-C, we provide the performance evaluation between the proposed method with other existing ones.

A. Evaluation Methodology

In the experiment, we adopt the eye tracking database [29] proposed in the study [23] to evaluate the performance of the proposed model. Currently, there are few available eye tracking database for 3D visual attention modeling in the research community. This database includes 18 stereoscopic images with various types such as outdoor scenes, indoor scenes, scenes including objects, scenes without any various object, etc. Some images in the database were collected from *the Middlebury 2005/2006 dataset* [42], while others were produced from videos recorded by using a Panasonic AG-3DA1 3D camera. To avoid the uncertainty from Depth of Field (DOF), the accommodation and vergence was considered within stereoscopic 3D viewing environment in this eye tracking experiment [29]. The disparity of the used stereoscopic images is within the comfortable viewing zone. Thus, the conflict from DOF will not detected by observers during this eye tracking experiments. However, DOF is normally associated with free vision in the real applications, where objects actually exist at different distances from observers. Some emerging stereoscopic displays are attempting to simulate this DOF effect in order to make the viewing experience more

TABLE I

Comparison Results of PLCC, KLD and AUC Values from Different Feature Maps: C_1 Feature Map: Color Feature Map from C_b Component; C_2 Feature Map: Color Feature Map from C_r Component; L Feature Map: Luminance Feature Map; T Feature Map: Texture Feature Map; D Feature Map: Depth Feature Map.

Models	PLCC	KLD	AUC
C_1 Feature Map	0.364	0.495	0.663
C_2 Feature Map	0.346	0.493	0.656
L Feature Map	0.438	0.449	0.672
T Feature Map	0.301	0.574	0.653
D Feature Map	0.443	0.4212	0.672
Final Saliency Map	0.703	0.260	0.740

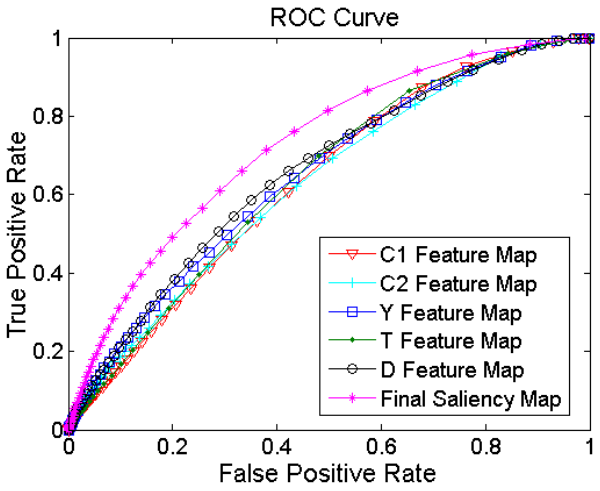
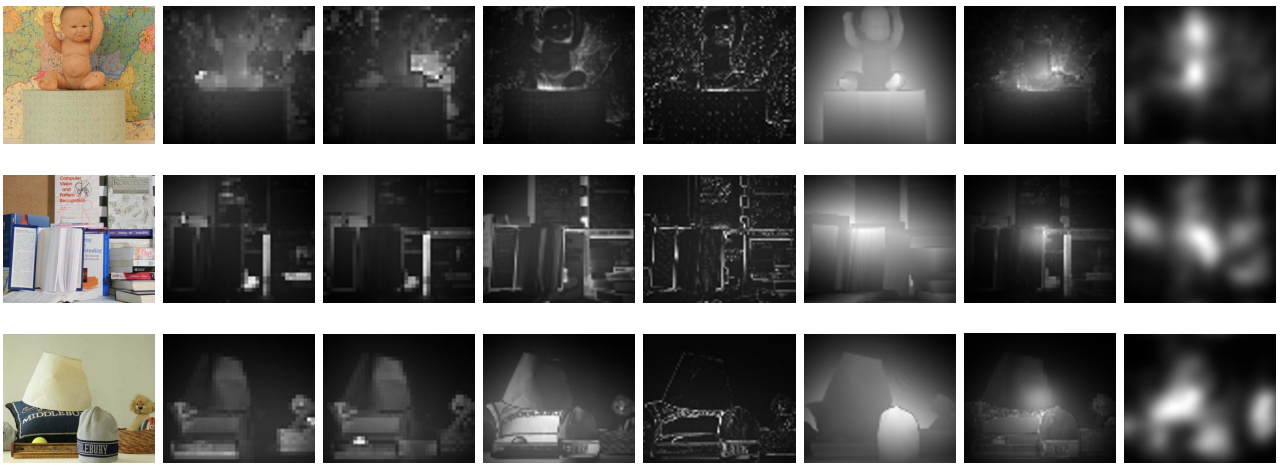


Fig. 4. The ROC curves of different feature maps: C_1 feature map: color feature map from C_b component; C_2 feature map: color feature map from C_r component; L feature map: luminance feature map; T feature map: texture feature map; D feature map: depth feature map.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. Visual comparison of saliency estimation from different features: (a) input image; (b) color feature map from C_b component; (c) color feature map from C_r component; (d) luminance feature map; (e) texture feature map; (f) depth feature map; (g) final saliency map; (h) ground truth map.

comfortable, especially for the case where 'near' objects exist. In the case where 'near' objects exist, there would be the narrowest DOF and thus it would attenuate the saliency of objects that are further away [51]. We will investigate more on the influence of DOF in stereoscopic saliency detection in the future work.

Stimuli were displayed on a 26-inch Panasonic BT-3DL2550 LCD screen with a resolution of 1920 × 1200 pixels and refresh rate of 60 Hz. The stereoscopic stimuli was viewed by participants with a pair o passive polarized glasses at a distance of 93 cm. The environment luminance was adjusted for each observer and thus the pupil had an appropriate size for eye-tracking. The data was collected by SMI RED 500 remote eye-tracker and a chin-rest was used to stabilize the observer's head. These gaze points recorded by eye-tracker are processed by a Gaussian kernel to generate the fixation density maps, which can be used as ground-truth maps. The images were presented in a random order and the presentation time for each image is 15 seconds. Thirty-five participants were involved in the eye tracking experiment. They ranged in age from 18 to 46 years old and the mean age is 24.2. All the participants had either normal or corrected-to-normal visual acuity, which was verified by pretests. Some samples of the left images and corresponding ground-truth maps are shown in the first and last columns of Fig. 6, respectively.

We use the similar quantitative measure methods as the study [23] for performance evaluation of the proposed method. The performance of the proposed model is measured by comparing the ground-truth and the saliency map from the saliency detection model. As there are left and right images for any stereoscopic image pair, we use the saliency result of the left image to do the comparison, similar with the study [23]. 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.

B. Experiment 1: Comparison Between Different Feature Channels

In this experiment, we compare the performance of different feature maps from color, luminance, texture and depth. Table I provides the quantitative comparison results for these feature maps. In this table, C_1 and C_2 color represent the color feature from Cb and Cr components respectively, which are described in Section III-A. From this table, we can see that the performance of saliency estimation from C_1 color feature is similar with that from C_2 color feature, while the feature map from Luminance feature can obtain better performance than that of color feature map from C_1 or C_2 component. Compared with color and luminance features, the depth feature can estimate better saliency result. For the texture feature, it gets the lowest PLCC and AUC values among these used features. Its KLD value is also higher than those from other features. Thus, the saliency estimation from texture feature is poorest among the used features. Compared with feature maps from these low-level features of color, luminance, texture and depth, the final saliency map calculated from the proposed fusion method can get much better performance for saliency estimation for 3D images, as shown by the PLCC, KLD and AUC values in Table I. The ROC curves in Fig. 4 also demonstrate the better performance of the final saliency map over other feature maps.

Fig. 5 provides some comparison samples of different feature maps and the final saliency map. From this figure,

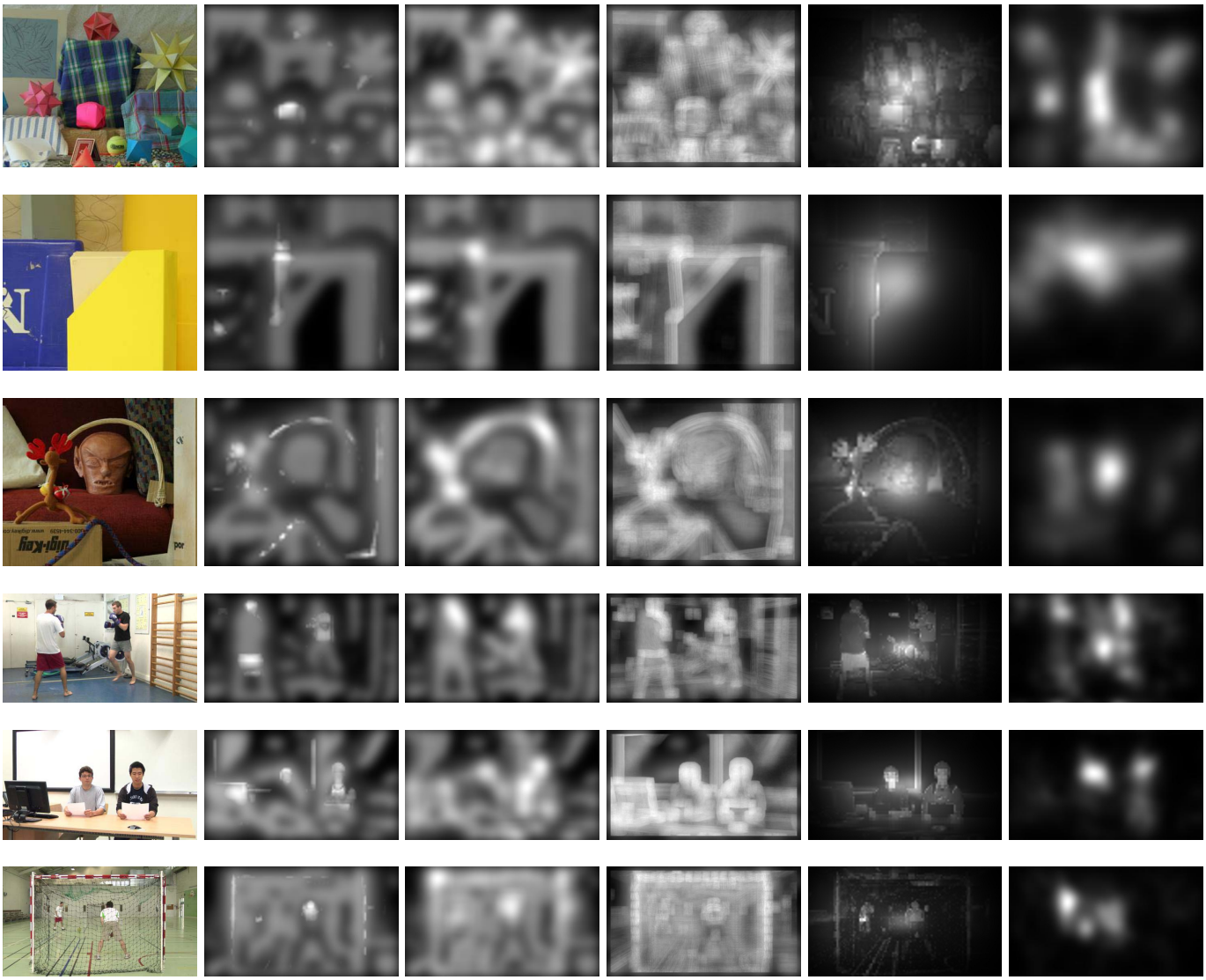


Fig. 6. Visual comparison of stereoscopic saliency detection models. (a) Input image. (b) Model 1 in [23]. (c) Model 2 in [23]. (d) Model 3 in [23]. (e) Proposed model. (f) Ground truth map.

we can see that the feature maps from color, luminance and depth are better than those from texture feature. The reason is that these features of color, luminance and depth represented by DC coefficients include much more energy for image patches compared with the texture feature represented by AC coefficients. Since AC coefficients only include high-frequency components in image patches, the feature maps from texture feature can mainly detect the shape of salient objects in images, as shown in Fig. 5(e). The overall saliency map by combining feature maps can obtain the best saliency estimation, as shown in Fig. 5(g).

C. Experiment 2: Comparison Between the Proposed Method and Other Existing Ones

In this experiment, we compare the proposed 3D saliency detection model with other existing ones in [23]. The quantitative comparison results are given in Table II.

TABLE II
COMPARISON RESULTS OF PLCC, KLD AND AUC VALUES FROM DIFFERENT STEREOSCOPIC-3D SALIENCY DETECTION MODELS. * MEANS THAT IT IS SIGNIFICANTLY DIFFERENT FROM THE PERFORMANCE OF THE PROPOSED MODEL (PAIRED T-TEST, $p < 0.05$)

Models	PLCC	KLD	AUC
Model 1 in [23]	0.356*	0.704*	0.656*
Model 2 in [23]	0.424*	0.617*	0.675*
Model 3 in [23]	0.410*	0.605*	0.670*
The Proposed Model	0.703	0.260	0.740

In Table II, Model 1 in [23] represents the 3D saliency detection model by fusion method of linear combination from 2D saliency detection model in [1] and depth model in [23];

TABLE III

CONTRIBUTION OF THE DEPTH INFORMATION ON 2D MODELS. + MEANS THE USE OF THE LINEAR POOLING STRATEGY INTRODUCED IN THE STUDY [23]. × MEANS THE WEIGHTING METHOD BASED ON MULTIPLICATION IN THE STUDY [23]. 2D REPRESENTS THE SALIENCY MAP FOR 2D IMAGES, WHILE DSM IS THE ABBREVIATION OF DEPTH SALIENCY MAP. * MEANS THAT IT IS SIGNIFICANTLY DIFFERENT FROM THE PERFORMANCE OF THE PROPOSED 3D FRAMEWORK (PAIRED T-TEST, $p < 0.05$)

		PLCC	KLD	AUC
IT (2D) [1]	2D Model Only	0.137*	2.819*	0.538*
	2D × DSM in [23]	0.137*	0.916*	0.540*
	2D + DSM in [23]	0.356*	0.704*	0.656*
	2D × Proposed DSM	0.242*	0.734*	0.671*
	2D + Proposed DSM	**0.448***	**0.416***	**0.676***
AIM (2D) [3]	2D Model Only	0.326*	0.736*	0.638*
	2D × DSM in [23]	0.403*	0.686*	0.671*
	2D + DSM in [23]	0.424*	0.617*	0.675*
	2D × Proposed DSM	0.486*	**0.384***	**0.688***
	2D + Proposed DSM	**0.491***	0.419*	0.683*
FT (2D) [4]	2D Model Only	0.291*	0.802*	0.630*
	2D × DSM in [23]	0.341*	0.782*	0.660*
	2D + DSM in [23]	0.410*	0.605*	0.670*
	2D × Proposed DSM	0.421*	0.495*	0.667*
	2D + Proposed DSM	**0.472***	**0.392***	**0.677***
Proposed Model	2D Model Only	0.496*	0.414*	0.682*
	2D × DSM in [23]	0.494*	0.485*	0.681*
	2D + DSM in [23]	0.534*	0.368*	0.694*
	2D × Proposed DSM	0.539*	0.336	0.704
	2D + Proposed DSM	0.549*	0.375	0.701
	The Proposed 3D Framework	**0.703**	**0.260**	**0.740**
Upper Theoretical Similarity Limit		0.897	0.127	0.782

Model 2 in [23] represents the 3D saliency detection model by fusion method of linear combination from 2D saliency detection model in [3] and depth model in [23]; Model 3 represents the saliency detection model by fusion method of linear combination from 2D saliency detection model in [4] and depth model in [23]. From this table, we can see that the PLCC and AUC values from the proposed model is larger than those from models in [23], while KLD value from the proposed model is lower than those from models in [23]. The statistical test results show the performance of the proposed model is significantly different from that from other existing ones. Thus, the proposed model can obtain a significantly higher performance than other existing models in [23]. The ROC curves in Fig. 7 also demonstrate the better performance of the proposed stereoscopic saliency detection model over other existing ones.

We also provide some visual comparison samples from different models in Fig. 6. From Fig. 6(b), we can see that the stereoscopic saliency maps from the fusion model by combining Itti's model [1] and depth saliency [23] mainly detect the contour of salient regions in images. The reason for this is that the 2D saliency detection model in [1] calculates saliency map mainly by local contrast. Similarly, there is the same drawback for the saliency maps from Fig. 6(c). For the saliency results from the fusion model by combing 2D saliency model in [3] and depth saliency in [23], some background regions are detected as salient regions in images, as shown in saliency maps from Fig. 6(d). 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. 6(e) and (f).

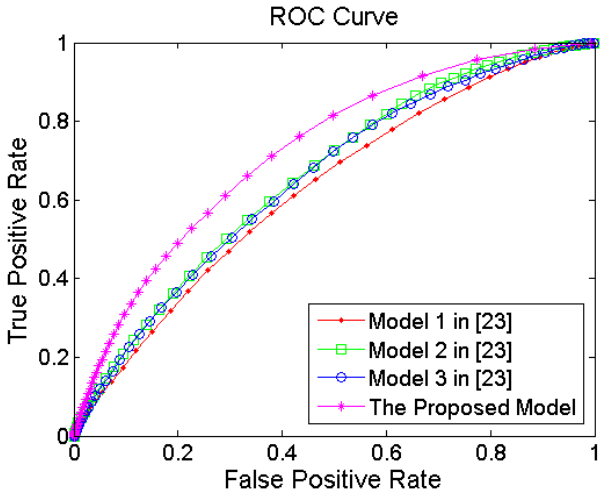


Fig. 7. The ROC curves of different stereoscopic saliency detection models.

To better demonstrate the advantages of the proposed algorithm, we compare the proposed algorithm and others from the aspects of 2D saliency and depth saliency in detail.

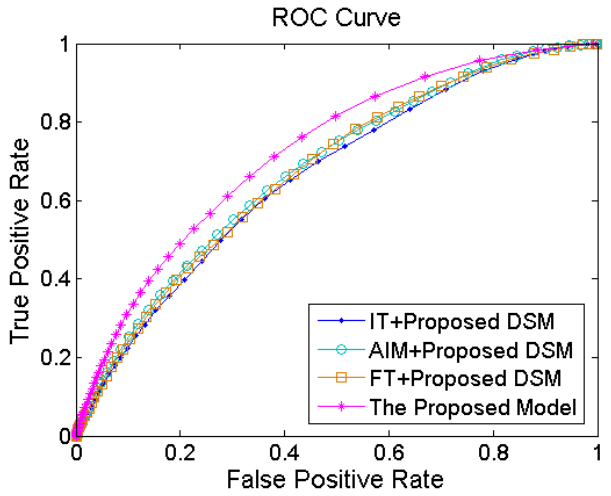


Fig. 8. The ROC curves of different stereoscopic saliency detection models. IT [1], AIM [3] and FT [4] are 2D saliency detection models.

Here, we use the same frameworks of depth-weighting combination method (the fusion method of multiplication combination) and depth-saliency combination method (the fusion method of linear combination) from different 2D and depth saliency maps to do the comparison. The quantitative comparison results and statistical test results are given in Table III. In the proposed model of Table III, we use the average combination for the feature maps from color, luminance, and texture features to obtain the proposed 2D model and combine it with the proposed DSM (Depth Saliency Map) to obtain the experimental results.

Table III provides the experimental results of 2D and 3D saliency detection models. From this table, we can see that the 3D saliency detection model with the depth information always obtains better performance than 2D saliency detection model, which demonstrates that the depth information is helpful in designing 3D saliency detection models. From the second row of Table III (IT (2D) [1]), we can see that the PLCC and AUC values from models by combing the 2D model and the proposed DSM are larger than those from models by combing the 2D model and DSM in [23], while the KLD values are smaller. This means that the saliency results from models by combing the 2D model and the proposed DSM are better than those from models by combing the 2D model and the DSM used in [23]. The third, fourth and fifth columns (AIM (2D) [3] and FT (2D) [4]) demonstrate the similar results. From these results, the 3D model by combing the 2D model and the proposed DSM can obtain better performance than others by combing the same 2D model and the DSM in [23]. Similarly, the 3D model by combing the proposed 2D and the DSM in [23] can get better performance than others by combing other 2D models and the DSM in [23]. From this table, the saliency results from the proposed 3D framework can get the significantly better performance than most of the compared models. We also provide the ROC curves of several compared models of Table III in Fig. 8. From this figure, we can see that the proposed model can obtain better performance than other compared ones.

TABLE IV
COMPARISON BETWEEN DIFFERENT 3D SALIENCY DETECTION MODELS. $\oplus$ MEANS THE COMBINATION BY SIMPLE SUMMATION INTRODUCED IN THE STUDY [26]. $\otimes$ MEANS THE COMBINATION BY POINT-WISE MULTIPLICATION IN THE STUDY [26]. DSM REPRESENTS THE DEPTH SALIENCY MAP FROM THE STUDY [26]. IT [1], GBVS [2], AIM [3], FT [4], ICL [47], LSK [48], AND LRR [49] ARE 2D SALIENCY DETECTION MODELS

Model	CC	AUC
IT $\oplus$ DSM	0.3752	0.8490
IT $\otimes$ DSM	0.3977	0.8539
GBVS $\oplus$ DSM	0.3903	0.8509
GBVS $\otimes$ DSM	**0.4128**	0.8546
AIM $\oplus$ DSM	0.3419	0.8495
AIM $\otimes$ DSM	0.3913	0.8503
FT $\oplus$ DSM	0.3148	0.7971
FT $\otimes$ DSM	0.2680	0.7449
ICL $\oplus$ DSM	0.3850	0.8455
ICL $\otimes$ DSM	0.3248	0.8077
LSK $\oplus$ DSM	0.3793	0.8453
LSK $\otimes$ DSM	0.3511	0.8237
LRR $\oplus$ DSM	0.3847	**0.8556**
LRR $\otimes$ DSM	0.3953	0.8463
The Proposed Model	**0.5453**	**0.8650**

Additionally, we use the recently published database from the study [26] to evaluate the performance of the proposed model. That database includes 600 stereoscopic images including indoor and outdoor scenes. These images are diverse with different objects, number and size of objects and degree of interaction or activity depicted in the scene. The eye tracker was used to record the human fixation from 80 participants. Here, we focus on the performance comparison between stereoscopic saliency detection models and use the fixation data from 3D images to conduct the experiment. Similar with the study [26], we calculate the AUC and CC (correlation coefficient) [50] values of the proposed model on the database. The experimental results are shown in Table IV. Please note that the AUC and CC values of other existing models are from the original paper [26]. From this table, we can see that the CC and AUC values from the proposed model are higher than other existing ones, which demonstrates that the proposed model can obtain better performance on saliency estimation on this database.

V. CONCLUSION

In this study, we propose a new stereoscopic saliency detection model for 3D images. 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. Additionally, we adopts the characteristics of the HVS (the centre bias factor and human visual acuity)

to enhance the saliency map. Experimental results show the promising performance of the proposed saliency detection model for stereoscopic images based on the recent eye tracking databases.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. NIPS*, 2006, pp. 545–552.

[3] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. NIPS*, 2006, pp. 155–162.

[4] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[5] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[6] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892–905, Aug. 2009.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[8] J. Yan, J. Liu, Y. Li, Z. Niu, and Y. Liu, "Visual saliency detection via rank-sparsity decomposition," in *Proc. IEEE 17th ICIP*, Sep. 2010, pp. 1089–1092.

[9] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1928–1942, Nov. 2005.

[10] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.

[11] Y. Fang, W. Lin, C. T. Lau, and B.-S. Lee, "A visual attention model combining top-down and bottom-up mechanisms for salient object detection," in *Proc. IEEE ICASSP*, May 2011, pp. 1293–1296.

[12] O. Le Meur, P. Le Callet, and D. Barba, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[13] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J. Visual Commun. Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[14] C. Guo and L. Zhang, "A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

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

[16] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

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

[18] N. Bruce and J. Tsotsos, "An attentional framework for stereo vision," in *Proc. 2nd IEEE Canadian Conf. Comput. Robot Vis.*, May 2005, pp. 88–95.

[19] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3d video," in *Proc. 16th Int. Conf. Adv. Multimedia Model.*, 2010, pp. 314–324.

[20] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, "Adaptive 3D rendering based on region-of-interest," *Proc. SPIE*, vol. 7524, Stereoscopic Displays and Applications XXI, 75240V, Feb. 2010.

[21] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *Proc. IEEE 15th Int. Conf. Pattern Recognit.*, Sep. 2000, pp. 375–378.

[22] E. Potapova, M. Zillich, and M. Vincze, "Learning what matters: Combining probabilistic models of 2D and 3D saliency cues," in *Proc. 8th Int. Comput. Vis. Syst.*, 2011, pp. 132–142.

[23] J. Wang, M. Perreira Da Silva, P. Le Callet, and V. Ricordel, "Computational model of stereoscopic 3D visual saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151–2165, Jun. 2013.

[24] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 454–461.

[25] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2011, pp. 409–416.

[26] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 101–115.

[27] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *Proc. BMVC*, 2013, pp. 112.1–112.11.

[28] L. Jansen, S. Onat, and P. Konig, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *J. Vis.*, vol. 9, no. 1, p. 29, 2009

[29] 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," *J. Eye Movement Res.*, vol. 5, no. 5, pp. 1–11, 2012.

[30] T. Jost, N. Ouerhani, R. V. Wartburg, R. Muri, and H. Hugli, "Contribution of depth to visual attention: Comparison of a computer model and human," in *Proc. Early Cognitive Vis. Workshop*, 2004, pp. 28.5–1.6.

[31] F. Shao, G. Jiang, M. Yu, K. Chen, and Y.-S. Ho, "Asymmetric coding of multi-view video plus depth based 3D video for view rendering," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 157–167, Feb. 2012.

[32] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1940–1953, May 2013.

[33] Q. Huynh-Thu, M. Barkowsky, and P. Le Callet, "The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 421–431, Jun. 2011.

[34] K. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Boston, MA, USA: Academic, 1990.

[35] C. Theoharatos, V. K. Pothos, N. A. Laskaris, G. Economou, and S. Fotopoulos, "Multivariate image similarity in the compressed domain using statistical graph matching," *Pattern Recognit.*, vol. 39, no. 10, pp. 1892–1904, 2006.

[36] B. A. Wandell. *Foundations of Vision*. Sunderland, MA, USA: Sinauer Associates, 1995.

[37] J. Wang, D. M. Chandler, and P. L. Callet, "Quantifying the relationship between visual salience and visual importance," *Proc. SPIE*, vol. 7527, Human Vision and Electronic Imaging XV, 75270K, Feb. 2010.

[38] W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," *Proc. SPIE*, vol. 3299, Human Vision and Electronic Imaging III, 294, Jul. 1998.

[39] J. Hakkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman, "What do people look at when they watch stereoscopic movies?" *Proc. SPIE*, vol. 7524, Stereoscopic Displays and Applications XXI, 75240E, Feb. 2010.

[40] C. Chamaret, O. Le Meur, and J. C. Chevet, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in *Proc. IEEE ICIP*, Sep. 2010, pp. 1077–1080.

[41] H. Nothdurft, "Salience from feature contrast: Additivity accross dimensions," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1183–1201, 2000.

[42] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2007, pp. 1–8.

[43] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–17, 2007.

[44] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, pp. 1–16, 2009.

[45] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, 2011.

[46] Y. Ma, L. Lu, H. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Int. Conf. Multimedia*, 2002, pp. 533–542.

[47] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. NIPS*, 2008, pp. 681–688.

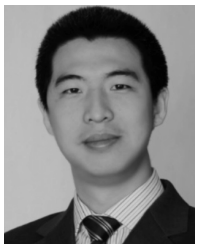
[48] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, 2009.

[49] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multitask sparsity pursuit," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1327–1338, Mar. 2012.

[50] N. Ouerhani, R. Wartburg, and H. Hugli, "Empirical validation of the saliency-based model of visual attention," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 3, no. 1, pp. 13–24, 2004.

[51] I. van der Linde, "Multi-resolution image compression using image foveation and simulated depth of field for stereoscopic displays," *Proc. SPIE*, vol. 5291, Stereoscopic Displays and Virtual Reality Systems XI, 71, May 2004.

Yuming Fang is currently a Lecturer with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China. He received the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2013. Previously, he received the B.E. and M.S. degrees from Sichuan University and the Beijing University of Technology, China, respectively. From 2011 to 2012, he was a Visiting Ph.D. Student with National Tsinghua University, Taiwan. From 2012 to 2012, he was a Visiting Scholar with the University of Waterloo, Canada. He was a (Visiting) Post-Doctoral Research Fellow with IRCCyN Laboratory, PolyTech' Nantes & Univ. Nantes, Nantes, France, the University of Waterloo, Waterloo, Canada, and the Nanyang Technological University, Singapore. His current research interests include visual attention modeling, visual quality assessment, image retargeting, and 3-D image/video processing. He was a Secretary of HHME2013 (the 9th Joint Conference on Harmonious Human Machine Environment), a Special Session Organizer for VCIP 2013, and a General Chair of the Third International Workshop on Emerging Multimedia Systems and Applications (in conjunction to ICME 2014).



Junle Wang received the Double M.S. degree in signal processing from the South China University of Technology, China, and in electronic engineering from the University of Nantes, France, in 2009, and the Ph.D. degree in computer science from the University of Nantes in 2012. He became an ATER (Assistant Professor) with the Department of Electronic and Digital Technologies, Ecole polytechnique de l'universié de Nantes. He is currently in charge of the Research and Development Division, Ars Nova Systems, France. His current research interests include image classification, visual attention, quality of experience of stereoscopic 3-D, image quality assessment, human visual perception, and psychophysical experimentation.

Manish Narwaria received the B.Tech. degree in electronics and communication engineering from Amrita Vishwa Vidyapeetham, Coimbatore, India, in 2008, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2012. He is currently a Post-Doctoral Researcher with IRCCyN-IVC Laboratory, France. His current research interests include signal processing and machine learning with applications in multimedia quality of experience, signal compression, and attention modeling.

Patrick Le Callet received the M.Sc. and Ph.D. degrees in image processing from Ecole polytechnique de l'Université de Nantes. He was a student with the Ecole Normale Superieure de Cachan where he sat the Aggrégation (credentialing exam) in electronics with the French National Education. He was an Assistant Professor from 1997 to 1999 and as a Full Time Lecturer from 1999 to 2003 with the Department of Electrical Engineering of Technical Institute of the University of Nantes (IUT). Since 2003, he has been teaching with Ecole polytechnique de l'Université de Nantes (Engineering School) in the Electrical Engineering and the Computer Science Departments where is currently a Full Professor. Since 2006, he has been the Head with the Image and Video Communication Laboratory, 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 research interests include 3-D image and video quality assessment, watermarking techniques, and visual attention modeling and applications. He is a co-author of more than 200 publications and communications and co-inventor of 13 international patents on these topics. He co-chairs within the Video Quality Expert Group, the Joint-Effort Group, and 3DTV activities. He is currently serving as Associate Editor for the IEEE TRANSACTIONS ON CIRCUIT SYSTEM AND VIDEO TECHNOLOGY, the *SPRINGER EURASIP Journal on Image and Video Processing*, and the SPIE Electronic Imaging.

Weisi Lin (M'92–SM'98) received the B.Sc. and M.Sc. degrees from Zhongshan University, China, and the Ph.D. degree from King's College, London University, U.K. He taught and researched in a number of organizations in China, U.K., and Singapore. He is currently the Associate Chair (Graduate Studies) in the School of Computer Engineering, Nanyang Technological University, Singapore. His areas of expertise include image processing, perceptual signal modeling, video compression and multimedia communication. He is a Chartered Engineer, a fellow of the IET, and an Honorary Fellow with the Singapore Institute of Engineering Technologists. He has been on the editorial boards of the IEEE TRANSACTION ON MULTIMEDIA (2011-2013), the IEEE Signal Processing Letters, and the *Journal of Visual Communication and Image Representation*. He holds seven patents, published more than 80 journal papers and 170 conference papers, authored a book, edited two books, and wrote seven book chapters; he has been elected as a Distinguished Lecturer of Asia-Pacific Signal and Information Processing Association from 2012 to 2013, and is an invited/panelist/keynote/tutorial speaker in VPQM06, SPIE VCIP10, the IEEE ICCCN07, PCM07, PCM09, the IEEE ISCAS08, the IEEE ICME09, APSIPA10, the IEEE ICIP10, the IEEE MMTC QoEIG in 2011, PCM 12, WOCC 12, HHME13, and APSIPA13. He guest-edited six special issues in international journals, and organized 12 special sessions in international conferences. He has served as a Technical Program Chair for PCM 12, ICME 13, and QoMEX 14.