Depth Matters: Influence of Depth Cues on Visual Saliency

Congyan Lang[1,3,⋆], Tam V. Nguyen[1,⋆], Harish Katti[2,⋆], Karthik Yadati[2], Mohan Kankanhalli[2], and Shuicheng Yan[1]

[1] Department of ECE, National University of Singapore, Singapore
{tamnguyen,eleyans}@nus.edu.sg
[2] School of Computing, National University of Singapore, Singapore
{harishk,nyadati,mohan}@comp.nus.edu.sg
[3] Department of Computer Science, Beijing Jiaotong University, China
cylang@bjtu.edu.cn

Abstract. Most previous studies on visual saliency have only focused on static or dynamic 2D scenes. Since the human visual system has evolved predominantly in natural three dimensional environments, it is important to study whether and how depth information influences visual saliency. In this work, we first collect a large human eye fixation database compiled from a pool of 600 2D-vs-3D image pairs viewed by 80 subjects, where the depth information is directly provided by the Kinect camera and the eye tracking data are captured in both 2D and 3D free-viewing experiments. We then analyze the major discrepancies between 2D and 3D human fixation data of the same scenes, which are further abstracted and modeled as novel depth priors. Finally, we evaluate the performances of state-of-the-art saliency detection models over 3D images, and propose solutions to enhance their performances by integrating the depth priors.

1 Introduction

Human visual exploration and selection of specific regions for detailed processing is permitted by the visual attention mechanism [1]. The eyes remain nearly stationary during fixation events as humans look at details in selected locations, which makes eye movements a valuable proxy to understand human attention. Visual saliency refers to the preferential fixation on conspicuous or meaningful regions in a scene [2] that have also been shown to correspond with important objects and their relationships [5]. Visual saliency is thus crucial in determining human visual experience and also relevant to many applications, such as automatic image collection browsing and image cropping.

Visual saliency has been extensively studied in signal processing, computer vision, machine learning, psychology and vision research literatures (e.g., [6,7,1,8,9]). However, most saliency models disregard the fact that human visual system operates in real 3D environments, while these models only investigate the cues from 2D images and the eye fixation data are captured in a 2D scene. However, stereoscopic contents provide additional depth cues that are used by

⋆ Indicates equal contribution.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 101–115, 2012.

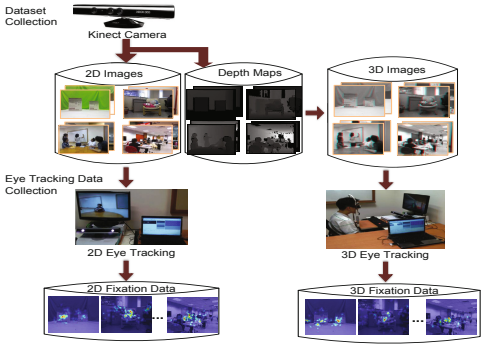


Fig. 1. Flowchart on 2D-vs-3D fixation dataset NUS-3DSaliency construction. We collect eye-tracking data on both 2D and 3D viewing settings and each 2D or 3D image was viewed by at least 14 observers. Eye fixations are recorded for each observer. The final fixation maps are generated by averaging locations across all the observers' fixations.

humans in the understanding of their surrounding and play an important role in visual attention [10]. Are the observer's fixations different when viewing 3D images compared to 2D images? How do the current state-of-the-art saliency models perform with additional depth cues? Not only are these questions interesting and important, answering them can also significantly benefit areas in computer vision research, such as autonomous mobile systems, 3D content surveillance and retrieval, advertising design, and adaptive image display on small devices.

In this paper, we conduct a comparative and systematic study of visual saliency in 2D and 3D scenes. Whereas existing eye tracking datasets captured for 2D images contain hundreds of images, the largest available eye tracking dataset for 3D scenes contains only a limited size of 28 stereo images [11]. A comprehensive eye tracking dataset for 3D scenes is yet to be developed. Motivated by these limitations, we collect a larger eye fixation dataset for 2D-vs-3D scenes. A 3D camera with active infra-red illumination (Microsoft "Kinect" [3]) offers the capability to easily extract scene depth information in order to extend 2D stimulus to 3D versions. Using an eye tracker, we collect eye fixation data to create human fixation maps which represent where viewers actually look in 2D and 3D versions of each scene. This large eye fixation dataset will be released for public usage. Our work further aims at quantitatively assessing the contribution of depth cues in visual attention in 3D scenes and proposing depth priors to improve the performances of state-of-the-art saliency detection models. In summary, the contributions of our work mainly include:

1. The eye-fixation dataset NUS-3DSaliency is collected from a pool of 600 images and 80 participants in both 2D and 3D scenes.
2. We analyze and quantify the difference between 2D and 3D eye fixation data. Based on the observations, the novel depth priors are proposed and integrated into saliency detection models.
3. We comprehensively evaluate the performances of state-of-the-art saliency detection models augmented with proposed depth priors on 2D and 3D eye fixation data.

2 Related Work

In order to understand what human attend to and qualitatively evaluate computational models, eye tracking data are used to create the human fixation maps, which will offer an excellent repository of ground truth for saliency model research. Most eye tracking datasets [7,12,13,5] are constructed for 2D scenes and most saliency models only investigate the cues from 2D images or videos. In contrast, relatively few studies have investigated visual attention modeling on 3D contents. Recently, several researchers have pioneered visual attention research on stereoscopic 3D contents. Jansen et al. [11] examined the influence of disparity on human behavior in visual inspection of 2D and 3D still images. They collected eye tracking data from 14 participants across 28 stereo images in a free-viewing task on 2D and 3D versions. A recent study [14] collected 21 video clips and the corresponding eye gaze data in both versions. However, compared with 2D eye tracking datasets, a comprehensive eye tracking dataset for 3D scenes is still absent. We believe that the studies on a rich and comprehensive 3D eye tracking dataset can offer interesting and important insights into how eye fixations are driven in real 3D environment. Motivated by this requirement, we collect a large 3D image database in this work, and then capture eye tracking data from an average of 14 participants per image across both 2D and 3D versions of 600 images.

Depth cues provide additional important information about contents in the visual field and can be regarded as relevant features for saliency detection [15] and action recognition [29]. Stereoscopic contents bring important additional binocular cues for enhancing human depth perception. Although there have been several efforts [16,17,18,19] to include the depth channel into computational attention models, a major problem in extracting depth from stereo input is the computation time needed to process disparities. In this paper, we study the discrepancies in human fixation data when viewing 2D and 3D scenes. The influence of depth on visual saliency is then studied and serves as the basis for learning depth priors to model 3D saliency.

3 Dataset Collection and Analysis

3.1 Dataset Collection

Our dataset aims to provide a comprehensive and diverse coverage of visual scenes for eye tracking analysis. We choose indoor and outdoor scenes that have natural co-occurrence of common objects. Furthermore, we systematically generate variants of scenes by varying parameters like depth ranges of different objects, number and size of objects and degree of interaction or activity depicted in the scene. We use the Kinect camera, which consists of an infra-red projector-camera pair as the depth camera that measures per pixel disparity, to capture a 640×480 pixel color image and its corresponding depth image at the same time. To the best of our knowledge, this is the largest 3D eye tracking dataset available to date for visual attention research, in terms of the total number of images and size of subject pool.

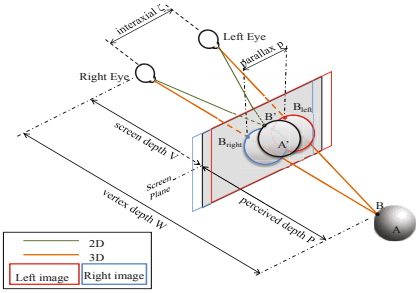


Fig. 2. The relationship between 2D and 3D fixations. The fixation location captured from participant viewing at B is the same for both 2D and 3D experiment setups. Screen depth W is the distance from the participant to the screen, while perceive depth P is calculated based on the depth value.

Stereoscopic Image Pair Generation for 3D Display. Following the collection of the color and depth image pair, the next step is to create 3D stimulus which involves generating left and right images. Prior to generating left-right image pair, some pre-processing on the captured images are required. We first perform calibration on both depth and color cameras to find the transformation between their images in a similar way as [4]. Next, we over-segment the color image into superpixels [20]. Each pixel, whose original depth value equal to 0, is assigned the average depth of the nearest neighbors in 8 directions in the same superpixel. Finally, we apply a default Laplacian filter with a 3×3 kernel for pixels whose depth values equal to 0 until all missing depth pixels are filled.

The stereoscopic image pair is produced by extracting parallax values from the smoothed depth map D and applying them to the left image I^l and right image I^r. For each pixel of the input color image I, the value of the parallax is obtained from its depth value. Figure 2 shows the relationship between 2D and 3D fixation. In both 2D and 3D viewing, for example, the fixation location for viewing B is recorded as the same position by the eye tracker. Considering the input image as a virtual central view, the left and right views are then obtained by shifting the input image pixels by a value ρ , $\rho = parallax/2$. In particular, the left image I^l and right image I^r can be obtained as $I^l(x_p^l) = I^r(x_p^r) = I(x_p)$, where x_p denotes the coordinate of the pixel in the color image I, the coordinate of the pixel in each view is calculated as $x_p^l = x_p + \rho, x_p^r = x_p - \rho$. Following Figure 2, the $parallax$ is calculated as follows:

$$parallax = \zeta \times (1 - \frac{V}{W}), \tag{1}$$

where ζ is the interaxial gap between two eyes, averaged as $60mm$, V is the screen depth or the distance from eyes to the screen and fixed as $80cm$ in our experiment setup, W is the vertex depth, equal to the summation of screen depth V and perceived depth P. For each pixel x in the image I, the perceived depth $P(x)$ can be calculated as $P(x) = D(x) \times \tau$, where τ ($\tau = 39.2$) is the ratio between the maximum depth distance captured by Kinect ($10{,}000mm$) and the

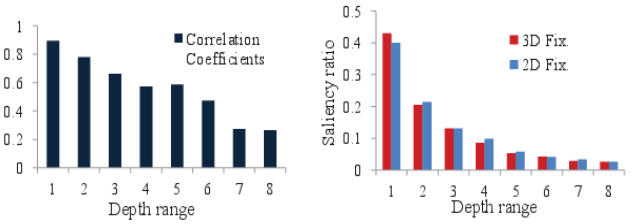
Fig. 3. Exemplar data in our eye fixation dataset NUS-3DSaliency. From left to right columns: color image and raw depth map captured by Kinect camera, smoothed depth map, 2D fixation map, and 3D fixation map.

maximum value in the depth image D (255). Since our dataset aims to provide the comprehensive and diverse coverage of visual scenes for eye tracking analysis, we reject images that are similar or have significantly overlapping content with other images in the dataset. Furthermore, images with significant artifacts after the smoothing process were rejected as an effort to minimize problematic images.

Participants. The participants (students and staff members of a university) ranged from 20 to 33 years old (μ=24.3, σ=3.1), among them 26 females and 54 males with normal or corrected-to-normal vision. All participants are naive to the purpose of the study and sign consent forms for public distribution of recorded eye-fixation data.

Data Collection Procedure. We use a block based design and free viewing paradigm. The subject views two blocks of 100 images that are unique and randomly chosen from the pool of 600 images, one of the blocks is entirely 2D and the other one entirely 3D. 3D images were viewed by using active shutter glasses on a 3D LCD display and 2D images were shown on the same screen in 2D display mode and the active shutter glasses switched off. In order to record subject eye gaze data, we used an infra-red illumination based remote eye-tracker from SensoMotoric Instruments Gmbh. The eye-tracker gives less than 1^o error on successful calibration. The eye tracker was calibrated for each participant using a 9-point calibration and validation method. Then images were presented in random order for 6 seconds followed by a gray mask for 3 seconds.

Human Fixation Maps. Human fixation maps are constructed from the fixations of viewers for 2D and 3D images to globally represent the spatial distribution of human fixations. Similar to [21], in order to produce a continuous fixation map of an image, we convolve a Gaussian filter across all corresponding viewers's fixation locations. Six examples of 2D and 3D fixation maps are shown



(a) (b)

Fig. 4. (a) The correlation coefficients between 2D and 3D fixations in different depth ranges. We observe lower correlation coefficients for farther depth ranges.(b) Saliency ratio in different depth ranges for 2D and 3D scenes respectively. The participants show to fix at closer depth ranges more often than farther depth ranges.

in Figure 3, the brighter pixels on the fixation maps denote the higher salience values.

3.2 Observations and Statistics

Using the recorded eye tracker data, we mainly investigate whether spatial distributions of fixations are different when human subjects view 3D images compared to 2D version. The interrelated observations are summarized as follows.

Observation 1: *Depth cues modulate visual saliency to a greater extent at farther depth ranges. Furthermore, humans fixate preferentially at closer depth ranges.*

In order to study the difference between 2D and 3D versions with respect to different depth range, fixation data for each 2D image I and its corresponding 3D image I', are divided into n ($n = 8$) depth ranges. Then for each depth range $r_b, b \in \{1, \cdots, n\}$, we compute the similarity between the 2D and 3D fixation distributions. We use the correlation coefficients (CC)[22] as similarity measure between two fixation maps. Figure 4(a) shows lower correlation coefficients for farther depth ranges.

Furthermore, in order to create a quantitative statistic of the relationship between the fixation distributions and depth ranges, we define saliency ratio as the description of fixation distribution. For the image I, we first compute saliency ratio $\gamma(r_b)$ as a function of the depth range, $\gamma(r_b) = \sum_x S(x)\delta(D(x) \in r_b)/\sum_x S(x)$, where $\delta(D(x) \in r_b)$ is set to 1 if x is in the depth range r_b. Figure 4(b) shows the saliency ratio vs. the depth range for 2D and 3D fixation data respectively. Looking at the data from the entire fixation dataset, the saliency ratio systematically decreases with the increase in depth range. A linear-best-fit of 2D(3D) saliency ratio with depth planes yields negative slopes $-0.045(-0.047)$ and indicates a statistically decreasing trend with increasing depth planes. From our analysis of fixation distribution and 2D-vs-3D correlation statistics over the entire dataset, we observe, (a) the larger discrepancy between 2D and 3D fixation data at further depth planes and, (b) the greater attenuation of visual attention at farther depth planes.

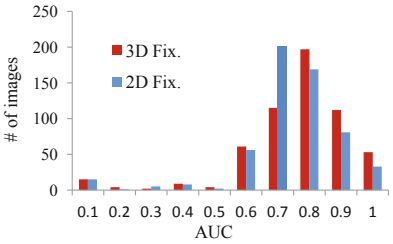


Fig. 5. We examine the ability of 2D/3D fixation map to predict the labeled interesting objects and histogram of the AUC values for 2D and 3D fixation dataset are comparatively shown is blue and red colors, respectively.

Observation 2: *A few interesting objects account for majority of the fixations and this behavior is consistent across both 2D and 3D.*

Such interesting objects such as human faces, body parts, text, cars and other conspicuous objects are discussed in [12]. Other studies such as [23] have shown that the eye fixations are correlated to the locations of such objects. In order to analyze this relationship, we follow the method in [12] by manually labeling objects of interest. To form more object-like contours, annotation of such regions is done by over-segmentation using superpixels [20] for each color image in our dataset. Despite occupying only 7.6% of the image area on average, the area corresponding to interesting objects account for 54.2% and 51.2% of the fixation points for 3D and 2D respectively. To quantitatively measure how well a 2D/3D fixation map predicts interesting objects on a given 2D/3D image, we compute the AUC value [8], the area under receiver operating characteristic (ROC) curve for each image. We use the labeled objects of interest as ground truth along with the fixation map as the predicted map of the image, this method effectively marginalizes out the influence of depth planes and helps to understand the role

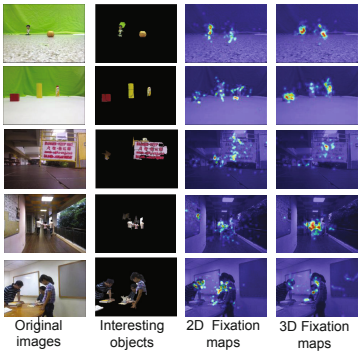


Fig. 6. Exemplar interesting objects manually labeled and fixation maps for 2D and 3D images. It indicates that the participants frequently fixed on such areas.

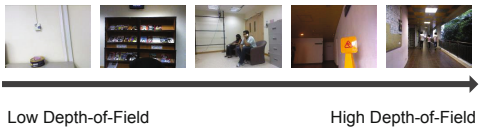


Fig. 7. Examples with low and high depth-of-field values

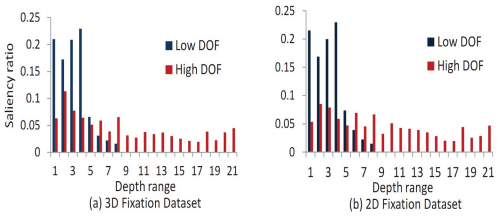


Fig. 8. Saliency ratio as a function of depth range. The saliency ratio distribution for 200 lowest depth-of-field images and for 200 highest depth-of-field images calculated on (a) 3D and (b) 2D fixation dataset respectively. The plot indicates that depth-of-field has influence on the allocation of attention in both 2D and 3D images.

of objects in isolation. Figure 5 shows the distribution of the AUC value for 600 labeled images. The average AUC for the entire 3D fixation dataset is 0.7399 and 0.7046 for 2D fixation dataset. Figure 6 gives examples of interesting objects along with corresponding fixation maps. These results suggest that the 2D and 3D fixation points show good correspondence with a few interesting objects.

Observation 3: *The relationship between depth and saliency is non-linear and characteristic for low and high depth-of-field scenes.*

Importantly, we observe that strong correlation exists between the depth-of-field (DOF) of the image and the fixation distribution. The DOF value ℓ of the image I is inferred from the distance between farthest and nearest depth values. In this experiment, we assign depth values into n ($n = 21$) depth ranges. The ℓ is defined as $\ell = |\overline{h^s} - \overline{h^t}|$, where $\overline{h^s}$ and $\overline{h^t}$ denotes the mean of the depth value for the pixels in the nearest and farthest depth ranges. Figure 7 shows some examples corresponding to the different DOF values. To demonstrate the influence of DOF, we analyze the saliency ratio defined in Observation 2 on two subsets of 200 images each selected from our dataset, one low-DOF subset and one high-DOF subset.The low DOF and high DOF partitions have a significant overlap of object types and this effectively marginalizes out the influence of objects. Similar to the statistic described in Observation 2, we create the statistic of the relationship between saliency ratio γ and the depth range for these two image subsets respectively. As shown in Figure 8, the saliency ratio distribution on different depth ranges have noticeable discrepancies between low DOF and high DOF images, as well as the distribution is non-linear. We find that 2D(3D) low DOF and corresponding 2D(3D) high DOF saliency ratio distributions in Figure 8 are dissimilar at $p = 0.05$ using a non parametric Kolmogorov-Smirnov test, on the other hand, saliency ratio distribution for 2D low(high) DOF shows

Table 1. The CC (correlation coefficient) comparison of fixation distribution on the 2D and 3D fixation data

DOF	0-0.25	0.25-0.5	0.5-0.75	0.75-1	*Avg.*
CC	0.8066	0.5495	0.2721	0.3057	0.4835

similarity to 3D low(high) DOF at $p = 0.05$. Motivated by this observation, we use the Gaussian Mixture Models (GMM) to model the distribution and the further implementation will be described in the next section.

Observation 4: *The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially, when there are multiple salient stimuli located in different depth planes.*

In order to study the difference between 2D and 3D versions, we divide the image dataset into four groups according the DOF values and compute the correlation coefficients between two fixation maps for the four groups respectively. Table 1 shows lower correlation coefficients for high DOF image groups. Figure 9 shows the fixation maps from the lower and higher depth-of-field images in 2D and 3D versions respectively.

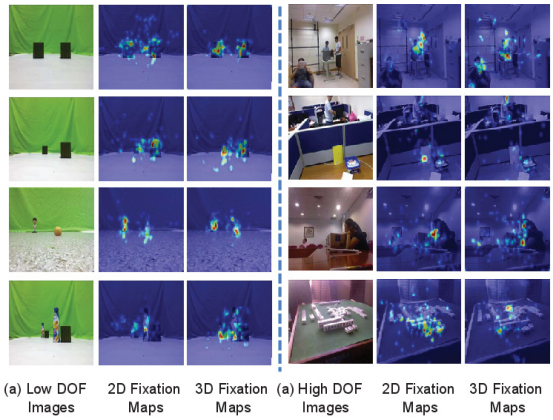


Fig. 9. Fixation maps and fixation distributions for 2D and 3D images. The results indicate a clear difference between 2D and 3D fixation maps with the increased Depth-of-field of the images.

4 Saliency Detection with Depth Priors

Based upon the above observations, we seek the global-context depth priors in order to improve the performance of the state-of-the-art saliency detection models. In this section, we propose to model the relationship between depth and saliency by approximating the joint density with a Mixture of Gaussians.

4.1 Learning Depth Priors

Formally, let D and S represent the depth image and fixation map of the image I, respectively. And d and s denote N-dimensional vector formed by orderly concatenating the patches from a regular partition of the image D and S respectively, N is the number of patches in the image. For the vector s, larger (smaller) magnitude implies that the patch is more salient (less salient). ℓ is the depth-of-field value introduced in the Section 3.2. The joint density between saliency response and depth distribution is written as

$$p(s, d|\ell) = \sum_{k=1}^{K} p(k)p(s|\ell, k)p(d|\ell, k), \tag{2}$$

where k indicates the kth component of the GMM. From the joint distribution we calculate the conditional density required for the depth modulated saliency:

$$\begin{aligned} p(s|d, \ell) &= \frac{p(s, d|\ell)}{\sum_{k=1}^{K} p(k)p(d|\ell, k)} \\ &\propto \sum_{c=1}^{C} q_c(\ell) \sum_{k=1}^{K} \pi_k \mathcal{N}(s; \mu_k^c, \Lambda_k^c)\mathcal{N}(d; \nu_k^c, \Upsilon_k^c), \end{aligned} \tag{3}$$

where $q_c(\cdot)$ is a quantization function for the depth-of-field. We compute a C-bins histogram of depth-of-field on the whole dataset. If ℓ falls into the cth bin, $q_c(\ell) = 1$, otherwise, $q_c(\ell) = 0$. Finally, the conditional expected saliency of the test image I_t, with the depth vector d_t and depth-of-field ℓ_t, is the weighted sum of K linear regressors:

$$\begin{aligned} s_t &= \sum_{c=1}^{C} q_c(\ell_t) \frac{\sum_{k=1}^{K} \mu_k^c * w_k^c}{\sum_{k=1}^{K} w_k^c}, \\ w_k^c &= \pi_k \mathcal{N}(d_t; \nu_k^c, \Upsilon_k^c). \end{aligned} \tag{4}$$

The model parameters are obtained from the training dataset and the EM algorithm is applied for fitting Gaussian mixtures. For the image I_t, its corresponding depth saliency map can be defined as the predicted saliency s_t. Note that s_t has a non-linear dependency with respect to the image depth distribution d_t.

4.2 Saliency Detection Augmented with Depth Priors

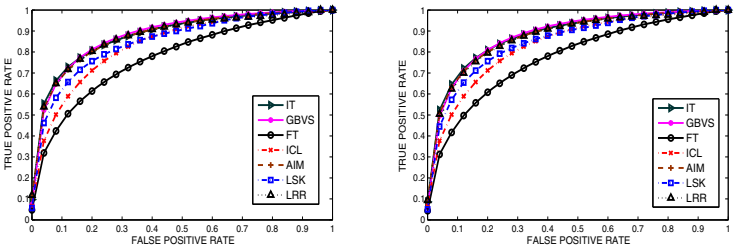
In order to investigate whether the depth priors are helpful for determining saliency, we extend seven existing methods to include the learned depth priors: Itti model (IT) [1], graph based visual saliency (GBVS) [8], frequency-tuned model (FT) [24], self-information (AIM) [7], incremental coding length (ICL) [25], local steering kernel (LSK) [26] and low-rank representation based model (LRR) [27]. The bottom-up saliency value predicted by the original models is

denoted as $\psi(x)$. The final saliency can be achieved[1] by simply using summation $\oplus$ or point-wise multiplication $\otimes$ as the fusion of two components. The final saliency is described by the equation:

$$S(x) = \psi(x)(\oplus/\otimes)p(s(x)|d(x), \ell). \tag{5}$$

5 Experiments and Results

In this section, we evaluate the saliency detection performance of the state-of-the-art models on our 2D and 3D fixation of NUS-3DSaliency dataset. Furthermore, we quantitatively assess the effectiveness of the depth priors improving the performances of saliency prediction algorithms.



(a) 2D fixation dataset (b) 3D fixation dataset

Fig. 10. ROC curves of different models. The results are from seven bottom-up saliency detection models by integrating depth priors to predict 2D and 3D fixation individually.

All saliency models use default parameter settings given by the corresponding authors. In order to learn depth priors, each image is resized to 200×200 pixels and regularly partitioned into 15×15 patches for training Gaussian models. The entire dataset is divided into four groups ($C = 4$) according to the depth-of-field value of each image. Depth saliency prediction is satisfactory with $K = 5$ as the number of the Gaussian components. For each image group, we randomly separate into 5 subsets, 4 subsets as the training set to learn the parameters of GMM and the remaining subset for testing. All the selected models are evaluated based on the following widely-used ROC and AUC. We also compute the correlation coefficients (CC) [22] between the fixation map and the predicted saliency map for evaluation.

5.1 Comparison of State-of-the-art Models

In this study, we examine the capability of seven bottom-up visual attention models to predict both the 2D and 3D fixation data in a free-viewing task.

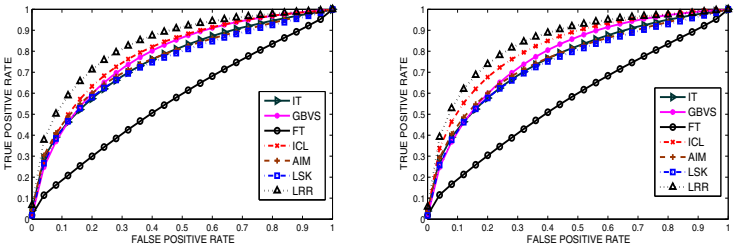
[1] Due to the space limitation, we do not further report how to more elegantly integrate depth priors into the models themselves, and only do simple late fusion in this work. We will explore the methods along this direction in our further work.

Table 2. The AUC and CC (correlation coefficient) comparison of different saliency models without the depth priors on the 2D and 3D eye fixation dataset

Criteria	AUC		CC	
Method	2D Fix.	3D Fix.	2D Fix.	3D Fix.
IT	0.7270	0.7299	0.2706	0.2594
GBVS	0.7486	0.7506	0.2986	0.2844
FT	0.5707	0.5726	0.1402	0.1388
ICL	0.7676	0.7673	0.3095	0.2759
AIM	0.7293	0.7308	0.2868	0.2531
LSK	0.7142	0.7158	0.2658	0.2425
LRR	**0.8045**	**0.7971**	**0.3155**	**0.2975**
2D Fix.	—	0.7982	—	0.3797
3D Fix.	0.8156	—	0.3797	—

Table 3. The AUC and CC (correlation coefficient) comparison of different saliency models with the depth priors on the 2D and 3D eye fixation dataset

Criteria	Method	2D Fix. ⊕	2D Fix. ⊗	3D Fix. ⊕	3D Fix. ⊗
AUC	IT	0.8521	0.8536	0.8490	0.8539
	GBVS	0.8541	**0.8562**	0.8509	**0.8546**
	FT	0.7995	0.7458	0.7971	0.7449
	AIM	0.8502	0.8517	0.8495	0.8503
	ICL	0.8406	0.8088	0.8455	0.8077
	LSK	0.8496	0.8233	0.8453	0.8237
	LRR	0.8511	0.8495	0.8556	0.8463
CC	IT	0.4000	0.4202	0.3752	0.3977
	GBVS	0.4128	**0.4346**	0.3903	**0.4128**
	FT	0.3355	0.2804	0.3148	0.2680
	AIM	0.3651	0.4180	0.3419	0.3913
	ICL	0.4126	0.3704	0.3850	0.3248
	LSK	0.4064	0.3764	0.3793	0.3511
	LRR	0.4065	0.4085	0.3847	0.3953

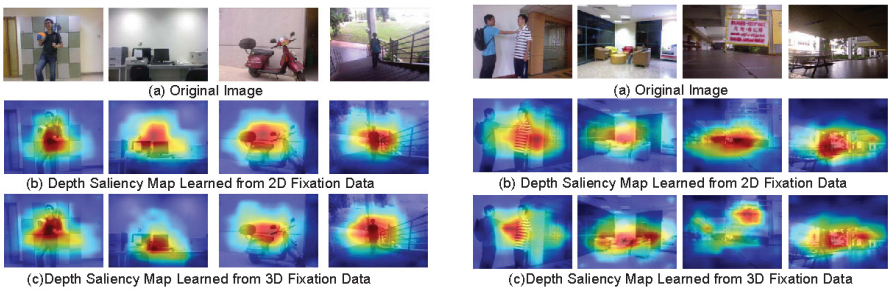


(a) 2D fixation data (b) 3D fixation data

Fig. 11. ROC curves of different models. The results are from seven bottom-up saliency detection models to predict on the 2D and 3D fixation data individually.

Figure 11 and Table 2 show the comparison results. First of all, most of models performed well for predicting human fixation when viewing 2D scenes. LRR exhibits stronger consistence with human eye fixations than the other models. We also evaluate the performance on 2D(3D) fixation maps to predict the 3D(2D) fixation maps. Interestingly, the AUC for the 2D fixation maps to predict 3D version is equal to 0.7982. On the contrary, the AUC of 0.8156 is obtained using 3D fixation maps to predict 2D fixation maps.

In contrast to 2D scenes, 3D scenes contain the additional information regarding the depth cue. This additional information could change the saliency of regions that are present in both 2D and 3D images. Here we show that the overall saliency is comparable in 2D and 3D scenes in terms of AUC. Thus, the bottom-up saliency models should also predict a fraction of the allocation of attention in 3D scenes. However, all models conducted on 2D scenes perform significantly better than 3D versions in terms of correlation coefficients. It further suggests that in a 3D attention model, depth could be considered as the important cue for saliency detection.



(a) (b)

Fig. 12. Representative examples in depth saliency prediction on 2D and 3D scenes respectively. (a) The predicted depth saliency maps are similar between 2D and 3D versions due to the scenes with one conspicuous area/object clearly standing out from the others. (b) The results show an obvious difference of the predicted depth saliency maps between 2D and 3D versions when multiply attractive objects or no conspicuous stimuli in the scenes.

5.2 Depth Priors for Augmented Saliency Prediction

In this subsection, we assess the influence of the depth priors on saliency detection. To evaluate quantitatively the effectiveness of the proposed depth priors, the results of saliency models integrating depth priors are shown in Table 3. The ROC curves are illustrated in Figure 10. These results show that the models with predicted depth priors perform consistently better than those without such depth priors. Overall we observe a 6% to 7% increase in predictive power using depth based cues. Another important aspect brought out in Table 2 a multiplicative modulating effect explains the influence of depth on saliency better than a linear weighted summation model, the latter has been used popularly in literature to combine results saliency maps derived from individual features [28].

Furthermore, Figure 12 gives some predicted saliency maps using depth priors alone (denoted as depth saliency map). On the one hand, the predicted depth saliency maps are similar in their spatial distribution between 2D and 3D versions when there is one conspicuous area or object clearly standing out from the others. On the other hand, when the scenes include multiple objects or no conspicuous objects, there is a noticeable difference between the predicted depth saliency maps of 2D and 3D cases.

6 Conclusions and Future Work

In this paper, we introduce the eye fixation dataset NUS-3DSaliency compiled from 600 images for both 2D and 3D scenes viewed by 80 participants. Using the state-of-the-art models for saliency detection, we have shown new performance bounds for this task. We expect that the newly built 3D eye fixation dataset would help the community advance the study of visual attention in a real 3D

environment. Furthermore, based on the analysis of the relationship between depth and saliency, extending the saliency models to include the proposed depth priors could consistently improve performance of the current saliency models. In future work, we are interested in how to integrate depth priors into various models instead of the current late fusion methods.

Acknowledgements. This work is partially supported by National Nature Science Foundation of China (90820013, 61033013, 61100142), Beijing Jiaotong University Science Foundation No. 2011JBM219; MOE project MOE2010-T2-1-087, Singapore; and the A*STAR PSF Grant No. 102-101-0029 on "Characterizing and Exploiting Human Visual Attention for Automated Image Understanding and Description".

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. TPAMI (1998)
2. Toet, A.: Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. TPAMI (2011)
3. Kinect: `http://www.xbox.com/kinect`
4. `http://nicolas.burrus.name/index.php/Research/Kinect/Calibration`
5. Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., Chua, T.-S.: An Eye Fixation Database for Saliency Detection in Images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 30–43. Springer, Heidelberg (2010)
6. Avraham, T., Lindenbaum, M.: Esaliency (extended saliency): Meaningful attention using stochastic image modeling. TPAMI (2010)
7. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: NIPS (2006)
8. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS (2006)
9. Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N.: Modelling visual attention via selective tuning. Artificial Intelligence (1995)
10. Wolfe, J., Horowitz, T.: What attributes guide the deployment of visual attention and how do they do it? Neuroscience (2004)
11. Jansen, L., Onat, S., Konig, P.: Influence of disparity on fixation and saccades in free viewing of natural scenes. JOV (2009)
12. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
13. Cerf, M., Frady, E., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. JOV (2010)
14. Quan, H., Schiaatti, L.: Examination of 3d visual attention in stereoscopic video content. SPIE-IST Electronic Imaging (2011)
15. Nakayama, K., Silverman, G.: Serial and parallel processing of visual feature conjunctions. Nature (1986)
16. Jang, Y.-M., Ban, S.-W., Lee, M.: Stereo Saliency Map Considering Affective Factors in a Dynamic Environment. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 1055–1064. Springer, Heidelberg (2008)

17. Aziz, M., Mertsching, B.: Pre-attentive detection of depth saliency using stereo vision. In: AIPR (2010)
18. Frintrop, S., Rome, E., Núchter, A., Surmann, H.: A bimodal laser-based attention system. CVIU (2005)
19. Ouerhani, N., Hugli, H.: Computing visual attention from scene depth. In: ICPR (2000)
20. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC Superpixels. EPFL Technical Report (2010)
21. Velichkovsky, B., Pomplun, M., Rieser, J., Ritter, H.: Eye-movement-based research paradigms. In: Visual Attention and Cognition (2009)
22. Ouerhani, N., Wartburg, R., Hugli, H.: Empirical validation of the saliency-based model of visual attention. Electronic Letters on CVIA (2004)
23. Meur, O., Chevet, J.: Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. TIP (2010)
24. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
25. Hou, X., Zhang, L.: Dynamic visual attention: searching for coding length increments. In: NIPS (2008)
26. Seo, H., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. JOV (2009)
27. Lang, C., Liu, G., Yu, J., Yan, S.: Saliency detection by multi-task sparsity pursuit. TIP (2011)
28. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. J. Electronic Imaging (2001)
29. Ni, B., Wang, G., Moulin, P.: HuDaAct: A color-depth video database for human daily activity recognition. In: ICCV Workshop (2011)