Extension of GBVS to 3D media

Zbigniew Zdziarski Rozenn Dahyot
zdziarze@tcd.ie Rozenn.Dahyot@tcd.ie
School of Computer Science and Statistics
Trinity College Dublin, Ireland

***Abstract*—Visual saliency has been studied extensively in the past decades through perceptual studies using eye tracking technologies and 2D displays. Visual saliency algorithms have been successfully developed to mimick the human ability to quickly spot informative local areas in images. This paper proposes to investigate the extension of visual saliency algorithms to media displayed in 3D. We show first that the Graph-Based Visual Saliency (GBVS) algorithm outperforms all the other common 2D algorithms as well as their 3D extensions. This paper then extends GBVS to 3D and shows that these new 3D GBVS based algorithms outperform other past algorithms.**

***Index Terms*—Visual saliency, 3D media.**

I. INTRODUCTION

Humans optimise the analysis of a visual scene by quickly sifting through irrelevant information. People's attention shifts from one *important* region to another and it is during this time that the brain builds its representation of the surrounding world [1]. This optimising behavioural action is the subject of research in visual saliency (VS) and it is studied extensively with the use of eye trackers monitoring people's gaze as they look at specific images and videos. In the past three decades, several visual saliency algorithms (VSAs) mimicking how people look at 2D images have been proposed [2]–[4] and have already proven to be valuable in several applications such as compression [5], content-based image retrieval [6], quality assessment [7] and retargetting [8].

Most of VS research has been focussing on the visual perception of images and videos displayed on 2D screens. More recently, however, it has been shown that humans look differently at images displayed in 3D compared to 2D [9], [10] and subsequently, a few algorithms for 3D saliency (3D-VSAs) have been proposed [11], [12]. The lack of ground truth has limited the comparison of VSAs in the past and recently Wang et al. [12] published an eye-tracking database with stereoscopic 3D images to act as ground truth to test VSAs. Using this database, they have proposed a comparison of several 2D-VSAs and 3D-VSAs [12]. We first propose to extend their study to include the Graph-Based Visual Saliency algorithm (GBVS) [2] and we show that GBVS performs better than other 2D-VSAs (Section IV). We then propose several extensions to GBVS to make it more effective for analysing 3D stereoscopic images (Section III). We show that our new algorithms (3D GBVS) outperforms other state of the art 3D-VSAs [12] (section IV). We start next with a review on VSAs and the metrics used to assess their performances.

978-1-4799-4874-1/14/$31.00

II. REVIEW

Visual saliency can be broken down into two groups: bottom-up and top-down [13] [14]. Bottom-up saliency is the preattentive, involuntary phase before any prior knowledge related to the task at hand or personal factors come into play that can affect attention guidance. Top-down saliency is highly dependent on semantic information and the current task being performed. For example, if we need to locate a red dragon in a box of toys, we would naturally scan the scene first for red objects and from these possibilities locate the desired toy.

This paper focuses on bottom-up saliency algorithms that are more generic (i.e. not dependent on the nature of the task) and faster to compute than top-down ones. Section II-A presents algorithms that do not incorporate any knowledge about the 3D nature of the scene depicted in the images (e.g. no depth or disparity is used to compute saliency). These algorithms are called here 2D-VSAs and Section II-B presents several extensions to these methods that include 3D information about the scenes (hence noted 3D-VSAs). Metrics have been introduced to measure the performance of VSAs using VS as ground truth captured with eye tracking technologies. Section II-C presents a few such metrics used in our experiments (Section IV).

A. 2D Visual saliency algorithms

Itti et al. [15] proposed an algorithm (noted 2D Itti) that first calculates features (i.e. colour, orientation and intensity) at multiple scales in an image. An activation map (an initial saliency map) is then calculated in parallel for each of these channels. Next, the activation maps are combined into a master saliency map. Finally, a ranking of salient regions is computed through the use of a Winner-Takes-All (WTA) network.

Hou et al.'s approach [16] (noted 2D Hou) analyses the log-spectrum of an image based on luminance only and extracts the spectral residual in the spectral domain. A saliency map is then created from this information. The algorithm is assessed in an object detection task and is shown to outperform Itti's method [16].

Bruce et al's algorithm [11] (noted 2D Bruce) is derived from efficient coding and information theory and is also shown to outperform Itti's algorithm. It is based on the premise that localised saliency computation serves to maximise information sampled from one's environment. Content of interest, hence, are areas where there is the most amount of 'surprise' and self-information.

Proposed by Harel et al. [2], the GBVS algorithm is a 2D VSA that uses a Markovian approach to calculate its saliency maps. The first step in this algorithm is to break up the input image into the following feature channels: colour, intensity and orientation. Salient regions are then located in each of these channels by computing:

$$d((i,j)||(p,q)) = \left|\log \frac{M(i,j)}{M(p,q)}\right| \quad (1)$$

where $M(i,j)$ is the value of the pixel (i,j) in the feature map M (i.e. in the feature channel of colour, intensity or orientation). Following this, a fully connected graph G_A is created by connecting every node with all other nodes in each M. The edge of each node connection from (i,j) to (p,q) is assigned the following weight:

$$w((i,j),(p,q)) = d((i,j)||(p,q)) \times F(i-p, j-p) \quad (2)$$

where

$$F(a,b) = \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right) \quad (3)$$

and σ is a free parameter set to approximatively one tenth to one fifth of the image width. All the weights, w, are, hence, proportional to their dissimilarity and distance in M. A Markov chain is then defined over G_A by treating nodes as states and edge weights as transition probabilities. The equilibrium distribution calculated for each G_A reflects the time a random walker would spend at each node. Higher values are obtained for nodes with higher dissimilarities with their neighbouring nodes. This is because it is more likely to transition into subgraphs with lower similarity measures. This algorithm has been stated as being biologically plausible, meaning that its method for calculating saliency is based on psychophysical and physiological evidence [2].

B. 3D Visual saliency algorithms

Stereoscopic images displayed on 3D screens allow us to immediately perceive depth information [17]. Incorporating depth or disparity information in the calculation of saliency is therefore a natural extension to VSAs for automatically analysing 3D content. Two strategies have been proposed to use depth information.

Chamaret et al. [18] have proposed to multiply the saliency map computed with any 2D VSAs by the inverse of the depth map. Regions appearing closer to the viewer are then made more salient. Indeed, close areas have been shown to be viewed more often than regions further away [10]. These regions, hence, should be deemed more salient. Chamaret et al. then used this saliency calculation to refine a single region-of-interest that was selected earlier through a nearest-neighbour filtering and thresholding approach.

Wang et al. [12] explore two common ways to include depth information. The first is the depth-weighting model that weights 2D saliency computations with a corresponding depth map. The second is the depth-saliency model that creates a depth saliency map (DSM) by first looking for features in the depth map and then linearly pooling this with 2D VSA computations. Wang et al.'s DSM calculations included a Difference of Gaussian (DoG) filtering step and then correlation of the contrast map with the degree of depth saliency through the use of results obtained from a psychophysical experiment of theirs. Through quantitative experiments, Wang et al. [12] show that the best algorithm for incorporing depth information was to add 2D VSAs saliency map with DSM. They also confirm quantitatively that 2D Bruce outperforms 2D Hou (and that 2D Hou outperforms 2D Itti) and this ordering remains true in 3D when adding the DSM. These algorithms proposed by Wang et al. are noted 3D Itti, 3D Hou and 3D Bruce. Adding DSM provided the best results but was the most computationaly demanding. Alternatively, Chamaret's method to include depth information performed poorly but was the most computationally efficient.

C. Metrics for VSAs

Several metrics have been introduced to measure the performance of VSAs. In our experiments, we used the Pearson Linear Correlation Coefficient (PLCC) [19], [20] and Kullback-Leibler divergence (KLD) [19], [21], which were also used in Wang et al.'s study [12]. These measures compare the saliency maps obtained from saliency algorithms with fixation density maps (maps created from eye-tracking experiments).

The PLCC measures the linear correlation between the saliency and fixation maps H and P:

$$\text{PLCC}(H,P) = \frac{\text{Cov}(H,P)}{\sigma_H \sigma_P} \quad (4)$$

where $\text{Cov}(H,P)$ is the covariance and σ_H and σ_P denote the standard deviations of H and P respectively.

The KLD calculates the dissimilarity between normalised saliency and fixation maps (to be understood as two probability density functions (PDFs)):

$$\text{KLD}(H,P) = \sum_x h_x \ln\left(\frac{h_x}{p_x}\right) \quad (5)$$

where h_x and p_x denote the values of the normalised maps of H and P respectively at pixel location x.

III. GBVS EXTENSIONS TO 3D MEDIA

The GBVS algorithm has never been extended to 3D and we propose to incorporate depth information in three ways: using Chamaret et al.'s approach [18] (noted 3D GBVS (Chamaret)), using the DSM as proposed by Wang et al. [12] (noted 3D GBVS (Wang)), and using our own approach (noted 3D GBVS (our approach)).

This last method (3D GBVS (our approach)) involves a three-step process: selecting a low or high scaling factor corresponding to depth values, restriction of the depth-range that saliency values are affected by this scaling factor and then the subsequent scaling of these saliency values. The first step entails calculating the median and mean values of the depth map. If the median is smaller than the mean, a higher scaling factor is chosen (SF in eqs 7 and 8) and vice versa otherwise. The idea behind this is to detect images that have

unique objects in the foreground and hence will stand out on their own in 3D. A smaller median value compared to the mean would provide an indication of this. This method is a fast way of detecting features in the depth map. The second step is performed by retaining only the depth values in the depth map that have a 2D saliency value over a chosen saliency threshold ST (saliency values are scaled between 0 and 1, where 1 indicates very salient and 0 indicates no saliency). We note these selected depth and 2D GBVS saliency values $\{(d_i, s_i)\}$ (i is the pixel index in the map) and find the range of these values by computing:

$$\begin{cases} l = \min_i\{d_i\} & \text{(lower bound)} \\ h = \max_i\{d_i\} & \text{(upper bound)} \end{cases} \quad (6)$$

and the middle of the interval is then defined by $dm = \frac{h+l}{2}$. For the pixel i, we compute

- If $d_i < dm$ then

$$S_i = \frac{s_i}{1 + SF \times \frac{dm-l}{dm-d_i}} \quad (7)$$

- If $d_i > dm$ then

$$S_i = s_i \left(1 + SF \times \frac{h - dm}{d_i - dm}\right) \quad (8)$$

where SF is the chosen scaling factor (high or low from step 1), S_i is the new 3D GBVS value, and s_i is the 2D GBVS value. We have set $SF = 0.35$ (high SF) or $SF = 0.1$ (low SF) and $ST = 0.4$ in our experiments. These threshold values were chosen after an exhaustive search of all possibilites. The justification behind the last two steps is that depth information should only be used on areas with high-enough saliency and should not be 'wasted' elsewhere. If salient regions are only present in the foreground, the competition for scaling should only take place there.

IV. Experimental results

We used the eye-tracking database supplied by Wang et al. [12] as ground truth. This database contains 18 stereoscopic images, eye-tracking data obtained from 35 human subjects, corresponding depth and disparity maps and eye fixation maps. This database serves as a ground truth and is noted 3D VS because the stereoscopic images were displayed on a 3D screen (as opposed to 2D VS ground truth that collects eye tracking data with the image displayed on a 2D screen).

A. Evaluation of 2D VSAs against 3D VS

In Table I, we compared the 2D GBVS algorithm on this dataset with Itti's, Hou's and Bruce's algorithms using the PLCC (defined in Eq. (4)) and KLD (Eq. (5)) metrics. Note that the 2D GBVS algorithm was not analysed by Wang et al. [12] but they did compute the PLCC and KLD for 2D Itti, 2D Hou and 2D Bruce. Our numerical results slightly differ from theirs due to rounding errors and different image scaling algorithms. Our results in Table I confirm Wang et al.'s assessment for 2D Itti, 2D Hou and 2D Bruce and our contribution in Table I is to show that 2D GBVS algorithm outperforms all other 2D VSAs by far.

	PLCC	KLD	Yr of publication
2D Itti	0.154	2.781	1998 [15]
2D Hou	0.299	0.877	2007 [16]
2D Bruce	0.346	0.704	2009 [11]
2D GBVS	**0.589**	**0.314**	**2007** [2]
UTPL [12]	0.897	0.127	

TABLE I: Performances of 2D VSAs. Higher PLCC and lower KLD values indicate better performances.

The Upper Theoretical Performance Limit (UTPL) [22] computed by Wang et al. [12] is also reported for both PLCC and KLD. The UTPL is commonly used as a benchmark for 2D visual saliency models, and 2D GBVS is halfway in performance between this theoretical limit and the best 2D VSA previously reported (2D Bruce). Figure 1 (c)-(f) shows some saliency maps for these 2D VSAs.

B. Evaluation of 3D VSAs against 3D VS

Our 3D VSAs are also evaluated on the same dataset. Wang et al. found their proposed 2D+DSM approach to be the best way to incorporate depth in saliency calculations, and this method is referred to as (Wang) in Table II. We implemented their approach for all the four 2D VSAs as well as Chamaret's and our own proposed method with the GBVS algorithm. Table II confirms the results for Itti (Wang), Bruce (Wang) and Hou (Wang) as reported by Wang et al. [12].

Our contribution in Table II is to show that the three 3D VSAs we proposed outperform all 3D VSAs proposed by Wang et al [12]. Surprisingly, Chamaret's method in the GBVS fared better than Wang's algorithm (that is also the most computationally demanding method). Note however that 3D GBVS (Chamaret) and 3D GBVS (Wang) are outperformed by 2D GBVS (cf. Table I). Our simple and fast proposed method for incorporating depth obtained the best results. Figure 1 (f)-(i) shows example saliency map results for 2D GBVS, 3D GBVS (Wang), 3D GBVS (Chamaret) and 3D GBVS (Our method). An improvement on 2D-GBVS can be seen in 3D GBVS (Chamaret) and 3D GBVS (Our method) - both are closer to the ground truth.

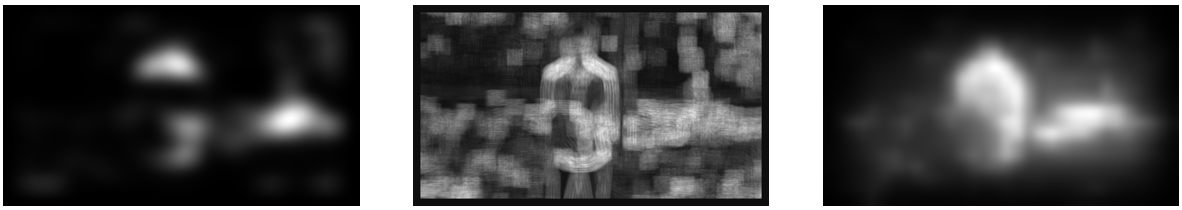
3D Saliency Algorithms	PLCC	KLD	Yr of pub.
3D GBVS (Chamaret)	0.573	0.379	
3D GBVS (Our method)	**0.606***	**0.306***	
3D GBVS (Wang)	0.561	0.484	
3D Itti (Wang)	0.364	0.627	2013 [12]
3D Bruce (Wang)	0.419	0.657	2013 [12]
3D Hou (Wang)	0.436	0.558	2013 [12]

TABLE II: Performances of 3D-VSAs. Higher PLCC and lower KLD values indicate better performance. * indicates significant difference from the 2D model (paired t-test, p<0.1 [12]).

V. Conclusion

In this paper we demonstrated that the GBVS algorithm is greatly superior in predicting fixations in 3D compared to other

(a) Original Image (b) Fixation density map (ground truth) (c) 2D Itti



(d) 2D Hou (e) 2D Bruce (f) 2D GBVS

(g) 3D GBVS (Wang) (h) 3D GBVS (Chamaret) (i) 3D GBVS (Our method)

Fig. 1: Saliency maps with VSAs. (a) Original image #18 [12], (b) Corresponding fixation density map, (c)-(f) Saliency predictions from the four 2D VSAs: Itti, Hou, Bruce and GBVS, (g)-(i) Saliency predictions from the three 3D VSAs: 3D GBVS (Wang), 3D GBVS (Chamaret) and 3D GBVS (Our method)

state of the art algorithms. We also showed that our simple and fast extension to Chamaret's method of depth incorporation outperforms all other depth incorporation methods analysed by Wang et al. Future work will investigate the application of our 3D VSAs in areas such as content-based image retrieval, compression and image segmentation.

REFERENCES

[1] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it," *Nature Reviews Neuroscience*, vol. 5, pp. 495–501, 2004.

[2] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proc. of Advances in Neural Information Processing Systems (NIPS)*, vol. 19, pp. 545–552, 2007.

[3] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," *Proc. of the Int'l Conf. on Pattern Recognition (ICPR)*, vol. 1, pp. 375–378, 2000.

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

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

[6] Z. Zdziarski and R. Dahyot, "Feature selection using visual saliency for content-based image retrieval," *IET Irish Signals and Systems Conf.*, 2012.

[7] Q. Ma and L. Zhang, "Image quality assessment with visual saliency," *Proc. of the Int'l Conf. on Pattern Recognition*, pp. 1–4, 2008.

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

[9] L. Jansen, S. Onat, and P. Konig, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *Journal of Vision*, vol. 9, no. 1, pp. 1–19, 2009.

[10] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel, and M. P. Da Silva, "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli," *Journal of Eye Movement Research*, vol. 5, no. 5, pp. 1–11, 2012.

[11] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[12] J. Wang, M. P. Da Silva, P. Le Callet, and V. Ricordel, "A computational model of stereoscopic 3d visual saliency," *IEEE Trans. on Image Processing*, vol. 22, no. 6, pp. 2151–2161, 2013.

[13] J. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2296–2303.

[14] S. Han and N. Vasconcelos, "Biologically plausible detection of amorphous objects in the wild," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 2011, pp. 24–31.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[16] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[17] K. Nakayama and G. H. Silverman, "Serial and parallel processing of visual feature conjunctions," *Nature*, vol. 320, pp. 264–265, 1986.

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

[19] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior research methods*, vol. 45, no. 1, pp. 251–266, 2013.

[20] A. Maeder and H. Zapernick, "Analysing inter-observer saliency variations in task-free viewing of natural images," *Image Processing (ICIP)*, pp. 1085–1088, 2010.

[21] O. Le Meur, P. Le Callet, D. Barba, and D. Thereau, "A coherent computational approach to model bottom-up visual attention," *Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[22] B. Stankiewicz, N. Anderson, and R. Moore, "Using performance efficiency for testing and optimization of visual attention models," *Proc. of SPIE*, vol. 7867, pp. 78670Y, 2011.