A low-complexity wavelet-based visual saliency model to predict fixations.

NARAYANASWAMY, M., ZHAO, Y., FUNG, W.K. and FOUGH, N.

2020

A Low-complexity Wavelet-based Visual Saliency Model to Predict Fixations

Manjula Narayanaswamy, Yafan Zhao, Wai Keung Fung and Nazila Fough
School of Engineering, Robert Gordon University, Aberdeen, United Kingdom, AB10 7GJ

***Abstract*— A low-complexity wavelet-based visual saliency model to predict the regions of human eye fixations in images using low-level features is proposed. Unlike the existing wavelet-based saliency detection models, the proposed model requires only two channels - luminance (Y) and chrominance (Cr) in YCbCr colour space for saliency computation. These two channels are decomposed to their lowest resolution using Discrete Wavelet Transform (DWT) to extract local contrast features at multiple scales. These features are integrated at multiple levels using 2D entropy based combination scheme to derive a combined map. The combined map is normalised and enhanced using natural logarithm transformation to derive a final saliency map. The experimental results show that the proposed model has achieved better prediction accuracy with significant complexity reduction compared to the existing benchmark models over two large public image datasets.**

Keywords—visual saliency model, fixation prediction, discrete wavelet transform, image entropy.

I. Introduction

The complexity of a visual scene is resolved by the human visual system by selectively attending the relevant regions of interest. The selective attention mechanism is achieved through a sequence of saccadic eye movements called fixations [2]. Predicting the regions of human eye fixations is essential where the identified regions can be used in the intelligent processing of visual information in computer vision systems. Visual saliency models are used to predict fixations in images and they have found vital importance in many areas such as image and video compression [3], image segmentation [5], remote sensing [7], and robotics [10].

Two types of attention mechanisms have been modelled in literature namely, bottom-up and top-down attention [11]. Predicting fixations based on bottom-up attention is a data-driven process which relies on sensory information of the input image [12], such as colour, luminance, motion, edges and so on. Whereas, predicting fixations based on top-down attention is a goal-oriented process which depends on high-level factors such as scene context, past knowledge and user expectations [11].

In literature, several computational visual saliency models with various approaches have been proposed. In 1998, Itti *et al.* [13] proposed a biologically inspired bottom-up model based on intensity, colour and orientation. Oliva *et al.* [14] proposed a probabilistic model with scene context as a top-down cue. Zhang *et al.* [15] proposed a SUN (Saliency using Natural Statistics) model which incorporates top-down information with bottom-up saliency, combined based on Bayes' rule. Hou *et al.* [16] proposed Spectral Residual (SR) model which uses amplitude spectrum of the Fourier transformed image to define saliency. In contrast to [16], Guo *et al.* [17] defined saliency using phase spectrum of the Fourier transformed image which was further extended to include motion features in [3]. Achanta *et al.* [18] proposed a frequency model, in which the difference of arithmetic mean vector of an image and its Gaussian blurred version is used to define image saliency. Chilukamari *et al*. proposed frequency models in [19, 20] to define saliency based on in-focus regions of an image.

The frequency models in [3, 16-18] are good at estimating image saliency on a global context but are inadequate as they do not contribute to local image saliency details. Conversely, computational models in Wavelet Transform (WT) domain [6, 8, 21, 22] have shown to be advantageous. This is mainly due to the property of WT that offers multi-resolution analysis which can be employed to perform local saliency analysis of an image at multiple scales. Murray *et al.* in [6] proposed a low-level vision model which uses biologically inspired Gabor-like wavelets to detect saliency at multiple scales. The model limitations occur from the centre-surround filtering process which makes it difficult to distinguish between true salient pixels and noisy pixels in the final saliency map. In contrast to [6], Imamoglu *et al.* [8] proposed a bottom-up model in the wavelet domain which incorporates global saliency along with local saliency information. The overall model out-performs the model in [6] but with high computational complexity involved. Ma *et al.* [22] proposed a bottom-up wavelet model to handle various contrast variations in different colour spaces. The limitations occur from Gaussian smoothing of saliency maps which causes heavy blurring and results in spatial information loss. Scharfenberger *et al.* [21] proposed a wavelet-based statistical model which defines saliency as non-redundant pixels at multiple scales. The model is mainly designed to deal with noisy images.

The aforementioned wavelet-based saliency detection methods have considerably achieved good prediction accuracy, but have not considered the complexity evaluation associated with the model. Many saliency applications deal with high-resolution images and real-time videos. It is a challenge to process excessive visual information with limited computational resources. Hence achieving good prediction accuracy while maintaining low computational complexity is critical for a saliency detection model and has a great impact on these applications in terms of efficiency and accuracy. The proposed model will focus on two objectives. Firstly, it aims to reduce the computational complexity of the model by limiting the number of colour channels required for saliency computation. Secondly, it aims to improve the prediction accuracy through an entropy-based feature combination scheme.

II. Proposed model

The model predicts the regions of human eye fixations in static images using the local contrast features of luminance (Y) and chrominance (Cr) channels of YCbCr colour space, combined using 2D entropy scheme and enhanced using natural logarithm transformation. Development of the proposed model consists of following four stages.

A. Colour transformation

The bright-light vision of human eye makes it more sensitive to brightness when compared to colour [12]. In the proposed work, the YCbCr colour space is preferred over RGB colour space as it can represent luminance (brightness) and chrominance channels separately. Thus, an input RGB image is converted to YCbCr colour space and convolved using a 2D Gaussian low-pass filter as in:

$$f(x,y) = I(x,y) * G_{s \times s} \tag{1}$$

where $I(x,y)$ is the input image channel with (x,y) being co-ordinates in 2D space, $G_{s \times s}$ is a 2D Gaussian low-pass filter with filter size s = 3, '*' is the convolution operator and $f(x,y)$ is the Gaussian smoothed channel. This filtering operation will eliminate very high-frequency noise present in the channels due to colour conversion.

B. *Multi-scale feature extraction*

The multi-resolution representation of the Discrete Wavelet Transform (DWT) provides an effective analysis of information content present in the images [23]. DWT uses a set of filters which decomposes the signal into independent frequency components (low-pass and high-pass). The local contrast variations are better represented in the high-pass frequency components of DWT which consists of details oriented in horizontal, vertical and diagonal directions at multiple scales [23]. Further, the experiments conducted by authors in [24] show that the relevant information can exist at different scales (from fine to coarse). Therefore, the local contrast features are extracted by decomposing the two effective channels Y and Cr (identified from the experimental results of TABLE I.) at N scales, where N is an integer given by $N = \log_2(D_{\max})$ with $D_{\max}$ being the maximum dimension of the input image. The biorthogonal wavelet 'bior4.4' with symmetrical nature of its wavelets and scaling functions, is chosen for decomposition [25]. In addition, it has also provided better saliency results when compared to other wavelets ('bior1.1', 'bior2.2' and 'bior3.3') of the family. Equation (2) represents DWT applied to the channel $f(x,y)$ at the i^{th} scale:

$$(f_i^a(x_a, y_a), f_i^s(x_s, y_s)) = DWT(f_i(x,y)) \tag{2}$$

where $f(x,y) \in$ {Y, Cr}, $i \in$ {1, 2, ..., N}, f^a represents transformed matrix that consists of low-frequency approximation (*a*) coefficients and f^s with $s \in \{h, v, d\}$ represents individual transformed matrices that consist of *h, v* and *d* coefficients respectively.

The Inverse DWT (IDWT) is applied to the high frequency *h, v* and *d* coefficients at *N* scales to derive feature maps at *N* levels as in [8]. This will create feature maps consisting of details from edge to texture at multiple decomposition levels. The IDWT operation is given in:

$$c_i(x,y) = IDWT(0, f_i^s(x_s, y_s)) \tag{3}$$

where $c_i(x,y)$ is the feature map obtained at the i^{th} level. The approximation details are omitted during reconstruction.

C. *Entropy based feature combination*

Entropy of an image can be defined as a statistical measure of information content present in the image [26]. In the proposed work, 2D entropy of a feature map is used as a weight to prioritise the feature combination. A feature map with high entropy value indicates a high saliency content and gets higher priority compared to the feature map with low entropy value. The feature map at each level is multiplied with its 2D entropy value as given in:

$$c'(x,y) = c(x,y) \times en_c \tag{4}$$

where $c'(x,y)$ is a weighted feature map, en_c is 2D entropy of feature map $c(x,y)$. The feature maps at N levels are combined as given in:

$$C(x,y) = \sum_{i=1}^{N} \left|y_i'(x,y)\right| + \left|cr_i'(x,y)\right| \tag{5}$$

where $\left|y'(x,y)\right|$ and $\left|cr'(x,y)\right|$ are the absolute values corresponding to respective Y and Cr weighted feature maps and $C(x,y)$ represents the combined map at N levels.

D. *Normalisation and enhancement*

The details in the combined map are smoothed using a 2D Gaussian low-pass filter with size 5. This will eliminate the high-frequency noise caused due to the wavelet processing. The combined map is normalised to a range [0, 1] which will ensure the details lie within the same range. Finally, the normalised map is enhanced using natural logarithm transformation to obtain the final saliency map. This will compress the dynamic range of intensity values to relatively a smaller intensity range [27] and enhances the salient regions in the final saliency map. Equation (6) represents the smoothing and normalisation operation:

$$C_M(x,y) = M(C(x,y) * G_{s \times s}) \tag{6}$$

where $M(.)$ denotes a normalisation operator, $C_M(x,y)$ denotes a normalised map and the filter size s = 5. Equation (7) represents the enhancement operation:

$$fmap(x,y) = \ln(C_M(x,y) + 1) \tag{7}$$

where ln(.) denotes natural logarithm transformation and $fmap\ (x,y)$ represents a final saliency map.

III. Experimental results

The proposed model is developed using MATLAB software. It is tested on two large public image datasets namely, MIT [1] and CAT2000 [4]. The image datasets consist of images under test and corresponding human ground truth (HGT) maps. The MIT dataset [1] consists of 1003 natural indoor and outdoor scenes. The corresponding HGTs were obtained from eye fixations of 15 observers. The

TABLE I. Experimental results for different channel combinations

Channel combination	**MIT dataset [1]**			**CAT2000 dataset [4]**		
	AUC	***CC***	***NSS***	***AUC***	***CC***	***NSS***
{Y, Cb, Cr}	0.70	0.24	0.82	0.70	0.31	0.78
{Y, Cb}	0.69	0.22	0.76	0.69	0.29	0.74
{Y, Cr}	**0.70**	**0.24**	**0.82**	**0.70**	**0.31**	**0.77**
{Y}	0.67	0.20	0.67	0.61	0.26	0.37

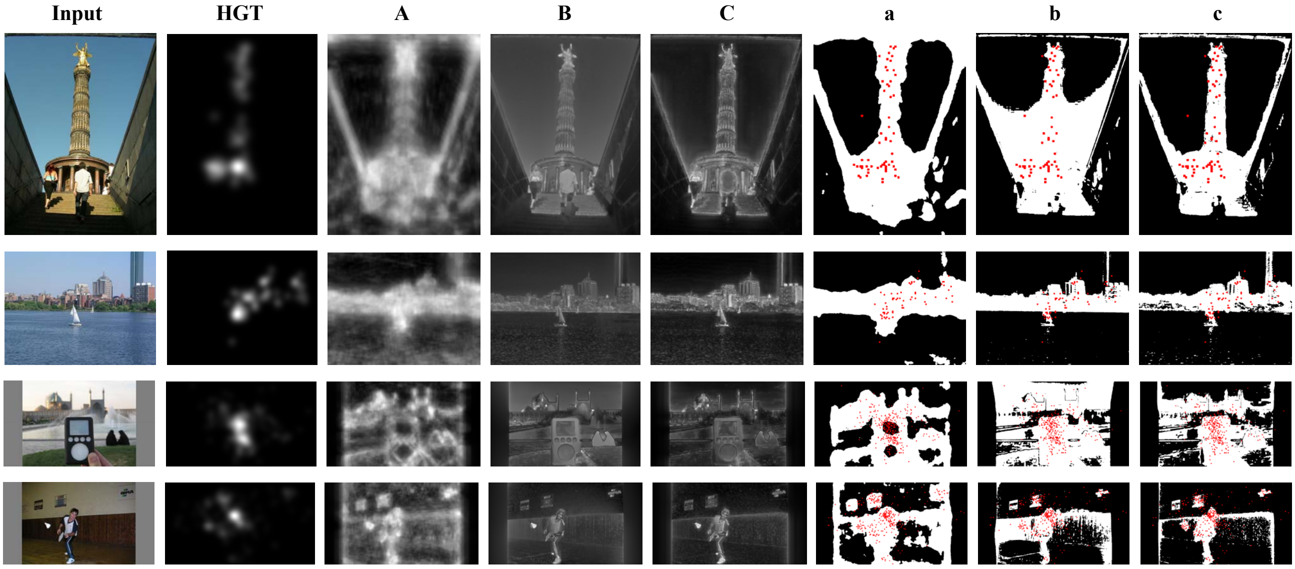

Fig. 1. Qualitative analysis of saliency maps. The input images of first two rows are obtained from MIT dataset [1] and the last two rows are obtained from CAT2000 dataset [4] respectively. Column A, B and C represents the saliency maps of Murray *et al.* [6], Imamoglu *et al.* [8] and Proposed model respectively. Columns a, b and c represents the corresponding threshold maps (obtained using Otsu's global threshold method [9]) with eye fixations (indicated by red markers).

CAT2000 [4] consists of 2000 natural and artificial images with corresponding HGTs obtained from eye fixations of 120 observers. The model is quantitatively evaluated using three performance metrics AUC (Area under ROC curve), CC (Pearson's Correlation Coefficient) and NSS (Normalised Scanpath Saliency).

To identify the most relevant channels required for saliency computation, the model scores were computed using different channel combinations. The experimental results provided in TABLE I. shows that better saliency scores are achieved for channel combinations {Y, Cb ,Cr} and {Y, Cr}. It is notable from the results of channel combination {Y, Cr} that channel Cb provides an insignificant contribution to saliency scores and can be eliminated. This achieves approximately one-third of complexity reduction in the bottleneck of the proposed model.

The prediction accuracy of the proposed model is evaluated qualitatively and quantitatively and compared with two state-of-the-art benchmark models, namely Murray *et al.* [6] and Imamoglu *et al.* [8]. Both the models detect saliency in images using wavelet coefficients which were developed in MATLAB with available open-source code. In order to obtain a fair comparison among the methods, the results for the model [8] are computed by ignoring the focus of attention concept. The qualitative results are shown in Fig. 1. The saliency maps provided in columns (A, B and C) indicate the grayscale information. The most salient pixels are represented with higher intensity values, in a decreasing order corresponding to the least salient pixels. The threshold or binary map provided in columns (a, b, and c) consists of salient regions (shown with white pixels) separated from background regions (shown with black pixels). The fixations in the first image coincide with the saliency maps of all the models. However, the saliency map of the proposed model has predicted reduced non-salient regions when compared to [6] and [8]. This can also be clearly observed in the fourth image. It can be seen from Fig. 1 that the saliency maps of the proposed model is able to detect true salient regions when compared to [6] which detects noise like salient regions. The results in the threshold maps show that the proposed method has provided a better correlation with fixations with reduced false detections when compared to [6] and [8].

The quantitative results are provided in TABLE II. The results show that the proposed model has outperformed in terms of CC and NSS with respect to [6] and [8], while provided similar or better performance in terms of AUC. The performance of the model is indicated using ROC (Receiver Operating Characteristics) curves as shown in Fig. 2 and 3. The higher portion of area under the curve (AUC) indicates better performance. Further, the computational complexity of the model has been evaluated over 100 images, randomly chosen from the MIT dataset [1] with resolution 768x1024 pixels. The results are compared with the models [6, 8] with the corresponding MATLAB code was obtained from online. The complexity evaluation results are provided in TABLE III. The test environment including, 16 GB RAM with a quad-core processor operating at a speed of 3.4GHz is used. It can be seen from the results in TABLE III. that the proposed model contributes to 91% of complexity reduction when compared to [8] and nearly 25% of complexity reduction when compared to [6].

TABLE II. QUANTITATIVE COMPARISON OF SALIENCY MODELS

Model	MIT dataset [1]			CAT2000 dataset [4]		
	AUC	*CC*	*NSS*	*AUC*	*CC*	*NSS*
Murray *et al.* [6]	0.70	0.23	0.78	0.69	0.28	0.72
Imamoglu *et al.* [8]	0.67	0.20	0.71	0.67	0.27	0.69
Proposed work	**0.70**	**0.24**	**0.83**	**0.70**	**0.31**	**0.77**

TABLE III. COMPLEXITY ANALYSIS OF SALIENCY MODELS

Model	Average Time (seconds)
Murray *et al.* [6]	2.45
Imamoglu *et al.* [8]	20.61
Proposed work	**1.84**

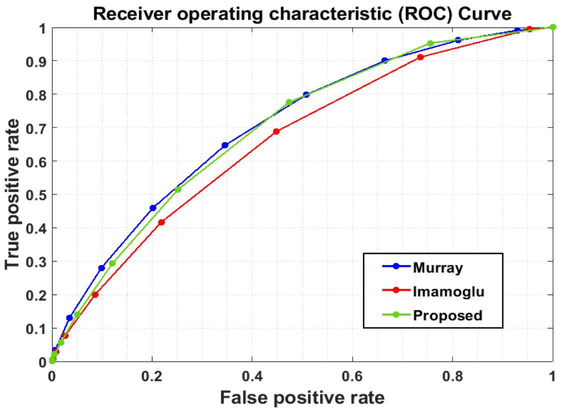

Fig. 2. ROC plot for MIT dataset [1]

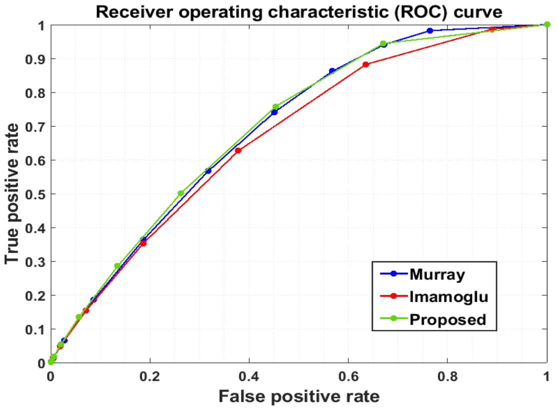

Fig. 3. ROC plot for CAT2000 dataset [4]

IV. Conclusion and Future Work

A low-complexity visual saliency model based on Wavelet Transform (WT) is proposed. The model predicts salient regions based on local contrast features of luminance and chrominance channels at multiple scales. Unlike the existing wavelet-based saliency detection methods, the proposed model requires only two-channel information for saliency computation. The experimental results show that our model has achieved significant complexity reduction (91% when compared to [8] and 25% when compared to [6]) and it has outperformed in terms of CC and NSS with similar or better performance in terms of AUC when compared to the models in [6] and [8].

The future work will focus on incorporating the global saliency information and top-down features for saliency detection in static images. Moreover, the temporal correlation and motion cues will be utilised to dynamically predict saliency in video sequences.

References

[1] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2106-2113.

[2] A. L. Yarbus, *Eye Movements and Vision*. New York: Plenum, 1967.

[3] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Transactions on Image Processing,* vol. 19, pp. 185-198, 2010.

[4] A. Borji and L. Itti, "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency," *Computer Vision and Pattern Recognition,* 2015.

[5] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient Object Segmentation via Effective Integration of Saliency and Objectness," *IEEE Transactions on Multimedia,* vol. 19, pp. 1742-1756, 2017.

[6] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Computer Vision and Pattern Recognition*, 2011, pp. 433-440.

[7] L. Zhang, Q. Sun, and Y. Sun, "Visual Saliency Analysis for Common Region of Interest Detection in Multiple Remote Sensing Images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2316-2320.

[8] N. Imamoglu, W. Lin, and Y. Fang, "A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform," *IEEE Transactions on Multimedia,* vol. 15, pp. 96-105, 2013.

[9] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 9, pp. 62-66, 1979.

[10] X. Yuan, J. Yue, and Y. Zhang, "RGB-D Saliency Detection: Dataset and Algorithm for Robot Vision," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2018, pp. 1028-1033.

[11] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 185-207, 2013.

[12] V. Bruce, P. R. Green, and M. A. Georgeson, *Visual Perception: Physiology, Psychology and Ecology*: Psychology Press, 2003.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1998.

[14] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, 2003, pp. I-253.

[15] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of vision,* vol. 8, pp. 32-32, 2008.

[16] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1-8.

[17] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[18] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597-1604.

[19] J. Chilukamari, S. Kannangara, and G. Maxwell, "A low complexity visual saliency model based on in-focus regions and centre sensitivity," in *2014 IEEE Fourth International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, 2014, pp. 411-414.

[20] J. Chilukamari, S. Kannangara, and G. Maxwell, "A DCT based in-focus visual saliency detection algorithm," in *2013 IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, 2013, pp. 1-5.

[21] C. Scharfenberger, A. Jain, A. Wong, and P. Fieguth, "Image saliency detection via multi-scale statistical non-redundancy modeling," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4294-4298.

[22] X. Ma, X. Xie, K.-M. Lam, and Y. Zhong, "Efficient saliency analysis based on wavelet transform and entropy," *Journal of Visual Communication and Image Representation,* vol. 30, pp. 201-207, 2015.

[23] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 11, pp. 674-693, 1989.

[24] L. Bonnar, F. Gosselin, and P. G. Schyns, "Understanding Dali's Slave Market with the Disappearing Bust of Voltaire: A case study in the scale information driving perception," 2001.

[25] D. L. Fugal, "Conceptual Wavelets in Digital Signal Processing," ed: Space & Signals Technical Publishing, 2009, pp. 5-5.

[26] N. D. B. Bruce and J. K. Tsotsos, "Saliency Based on Information Maximization " presented at the Neural Information Processing Systems, 2005.

[27] R. C. Gonzalez and R. E. Woods. (2018). *Digital Image Processing (Fourth ed.).*