Bottom-Up Visual Saliency Using Binary Spectrum of Walsh-Hadamard Transform

Ying Yu[1], Jie Lin[2], and Jian Yang[1]

[1] School of Information Science and Engineering, Yunnan University, Kunming 650091, China
[2] Department of Information Management, Yunnan Normal University, Kunming 650500, China
yuying.mail@163.com, linjie@ynnu.edu.cn, nxryang@126.com

Abstract. Detection of visual saliency is valuable for applications like robot navigation, adaptive image compression, and object recognition. In this paper, we propose a fast frequency domain visual saliency method by use of the binary spectrum of Walsh-Hadamard transform (WHT). The method achieves saliency detection by simply exploiting the WHT components of the scene under view. Unlike space domain-based approaches, our method performs the cortical center-surround suppression in frequency domain and thus has implicit biological plausibility. By virtue of simplicity and speed of the WHT, the proposed method is very simple and fast in computation, and outperforms existing state-of-the-art saliency detection methods, when evaluated by using the capability of eye fixation prediction.

Keywords: Visual attention, Saliency detection, Walsh-Hadamard transform.

1 Introduction

Visual saliency refers to the perceptual quality that makes an object or location stand out or pop out relative to its neighbors and thereby attract visual attention. Typically, visual attention is either driven by fast, pre-attentive, bottom-up visual saliency, or controlled by slow, task-dependent, top-down cues [1].

This paper is primarily concerned with the automatic detection of bottom-up visual saliency, which has already attracted intensive investigations in the area of computer vision in relation to robotics, cognitive science and neuroscience. One of the most influential computational models of bottom-up saliency detection was proposed by Itti et al. [2], which is designed conforming to the neural architecture of the human early visual system and thereby has biological plausibility. Itti et al.'s model has been shown to be successful in detecting salient objects and predicting human fixations. However, the model is ad-hoc designed and suffers from over-parameterization.

Some recent works addressed the question of "what attracts human visual attention" in an information theoretic way, and proposed a series of attention models based on information theory. These models based on information theory include the attention model based on information maximization [3], the graph-based visual saliency approach [4], and the discriminant center-surround approach [5]. While these

C.K. Loo et al. (Eds.): ICONIP 2014, Part III, LNCS 8836, pp. 33–41, 2014.

information theory-based models show better performance in saliency detection than Itti et al.'s model, they are more computationally expensive for some real-world systems.

Another kind of saliency models are implemented in the frequency domain, which are not at all biologically motivated, but they have fast computational speed and good consistency with psychophysics. These frequency domain models include the so-called spectral residual approach [6], and the approach using phase spectrum of quaternion Fourier transform [7]. Later works proposed by Yu et al. [8][9] asserted that visual saliency can be describes in terms of spatial correlation in the visual space, and that saliency information can be generated within a simple normalization process for principal component analysis (PCA) coefficients of the scene under view. Yu et al.'s saliency model has neurobiological plausibilities because the principal components of natural scenes can be obtained by using a Hebbian-based neural network.

In this paper, we propose a bottom-up visual saliency method based on the Walsh-Hadamard transform (WHT). Our saliency method simply projects the whole image into the WHT space and utilizes the signs of the WHT components to compute the saliency information of the visual space. This significantly reduces computations because unlike all spatial domain approaches, our method does not need to decompose the input image into numerous feature maps separated in orientation and scale, and then compute saliency at every spatial location of every feature map. Such a computation process may be quick for the massively parallel connections of the human visual pathway, but is comparatively slow for computer processors. The WHT [10][11] is perhaps the most well-known of the non-sinusoidal orthogonal transforms, which has gained prominence in various digital signal processing applications, since it can essentially be computed using additions and subtractions only. Consequently its hardware implementation is also simpler. The proposed saliency method is referred to as binary spectrum of Walsh-Hadamard transform (BWHT) in this paper. As compared to other frequency domain approaches, our method is simpler and faster in computation, and requires fewer storage spaces.

The remainder of this paper is organized as follows. Section 2 describes the proposed method of bottom-up visual saliency as well as its neurobiological plausibility. Section 3 presents the experiments and quantifies the consistency of our saliency method with eye fixation data. Finally, conclusions are given in Section 4.

2 Proposed Method

In this section, we begin by providing an interpretation of bottom-up visual saliency, and then propose a saliency detection method based on the WHT. We will explain how our proposed method relates to visual saliency.

2.1 Visual Saliency

Li [12] hypothesized that the primary visual cortex (V1) creates a bottom-up saliency map of the visual space and the contextual influence is necessary for saliency computation. For example, a red flower is salient in a context of green leaves. Each neuron in V1 is tuned to a particular visual feature such as color and orientation. The dominant contextual influence in V1 is the so-called "iso-feature suppression", i.e., nearby neurons tuned to similar features are linked by intra-cortical inhibitory connections [13]. Besides Li's hypothesis, a number of recent studies (e.g., [3][5][14][15]) have attempted to describe visual saliency in terms of surprise, interest, innovation, self-information and center-surround discrimination. These studies provided a general idea that higher information entropy accounts for higher saliency.

Our visual environment is highly structured and thereby much information redundancy exists in the visual input. It has been shown that the dominant redundancy of our visual input arises from second order input statistics and that the human visual system is capable of reducing such redundancy of visual sensory data [16]. Yu et al. [9] found that visual saliency can be described in terms of statistical correlation in the visual space, and employed the PCA projection vectors to capture the second order correlated components among image pixels. They have attempted to suppress highly correlated image components and meanwhile highlight salient image regions by normalizing the PCA coefficients of the input image. Following Yu et al.'s interpretations of visual saliency, in the next subsection we use the WHT to capture highly correlated components in visual space and suppress them so as to highlight salient visual features.

2.2 Saliency Map

It has been noted that like the PCA for natural images, the WHT components reflect global features in the visual space, and the redundancy reflected in the second-order correlations between pixels can be captured by the WHT components of the image [10][11]. According to such an interpretation of visual saliency in the previous subsection, image regions with high spatial correlation with its surroundings can be suppressed through a normalization operation upon the WHT components. As a result, salient locations can be relatively highlighted.

As compared to the PCA for natural images, the WHT is much simpler and faster, and has many fast algorithms for its computation. Moreover, a 2-dimensional WHT is separately performed in row and column, and therefore its computational complexity is significantly lower than a PCA transformation.

We start by considering a gray-scale image $\boldsymbol{X}$. According to previous analysis, we first conduct a 2-dimensional WHT on the image. Next, we normalize the WHT components by setting all positive coefficients to a value of 1 and all negative coefficients to a value of -1. This 2-dimensional orthogonal transformation followed by a normalization operation can be easily formulated as

$$\boldsymbol{B} = \mathrm{sign}(\mathrm{WHT}(\boldsymbol{X})), \tag{1}$$

where "$\mathrm{WHT}(\cdot)$" denotes a 2-dimensional Walsh-Hadamard transform, and the notation "$\mathrm{sign}(\cdot)$" is a signum function. The matrix $\boldsymbol{B}$ is referred to as binary spectrum of Walsh-Hadamard transform (BWHT) in this paper. It retains only the sign of each WHT component, discarding the amplitude information across the entire frequency spectrum. Note that $\boldsymbol{B}$ is expressed in binary codes (i.e., 1s and -1s) and thereby is very compact, with a single bit per component. The signum function, which normalizes the WHT coefficients, suppresses highly correlated components in the visual space and thereby accomplishes the computation of visual saliency in the WHT domain.

To recover the saliency information in the visual space, we conduct an inverse WHT on the binary spectrum $\boldsymbol{B}$, which is formulated as

$$\boldsymbol{F} = \mathrm{abs}(\mathrm{IWHT}(\boldsymbol{B})), \tag{2}$$

where "$\mathrm{IWHT}(\cdot)$" denotes the corresponding inverse Walsh-Hadamard transform, and the notation "$\mathrm{abs}(\cdot)$" is an absolute value function. Normally, the obtained matrix $\boldsymbol{F}$, which carries the saliency information, is post-processed by convolution with a Gaussian filter for smoothing. This operation can be formulated as

$$\boldsymbol{S} = \boldsymbol{G} * \boldsymbol{F}^2, \tag{3}$$

where $\boldsymbol{G}$ is a 2-dimensional Gaussian kernel, and $\boldsymbol{S}$ is the corresponding saliency map of the input image $\boldsymbol{X}$. Note that $\boldsymbol{F}$ is squared for visibility.

It is worth stating that we resize the image to a width of 64px and keep its aspect ratio before computing the saliency map. This spatial scale is chosen according to the heuristics of other frequency domain approaches (e.g., [6][7][9]).

In the human visual pathway, the color space of natural images is decomposed into well decorrelated channels. The RGB color space is highly correlated, but an LAB color space transformation results in well decorrelated color channels for natural color images. In addition, the transformation is perceptually uniform, and it produces three biologically plausible channels: a luminance channel, a red-green opponent channel and a blue-yellow opponent channel.

The complete BWHT algorithm from input image to final saliency map is given as follows.

1. Perform an LAB color space transformation
2. Resize the image to a suitable scale
3. Perform a Walsh-Hadamard transform for each color channel and calculate the binary spectrum of all WHT components using equation (1)
4. Obtain the saliency maps of each color channel using equation (2)
5. Take the spatial maximum across the saliency maps of all color channels to obtain the final saliency map
6. Post-process the saliency map by convolution with a Gaussian filter for smoothing and visibility as formulated in equation (3)

For recombination, we take the maximum value, as argued by Li and Dayan [13], at each pixel location of the corresponding saliency maps instead of spatial summation used by most models. The complete flow of the proposed method is illustrated in Fig. 1. The input image is initially decomposed into three biologically motivated channels: a luminance channel and two color opponent channels. Each of the three channels is then subjected to a Walsh-Hadamard transformation. Then, the binary spectrum of WHT is obtained by taking the signs of the WHT components of each channel. Afterward, the binary spectrum of each channel is subjected to an inverse Walsh-Hadamard transformation so that the saliency map of each channel is generated. Finally, a final saliency map is obtained by taking the spatial maximum value across all three saliency maps. Note that the saliency map is a topographically arranged map that represents visual saliency of a corresponding visual scene. The objects or locations with high saliency values may stand out or pop out relative to their surroundings, and thus attract our visual attention. From Fig. 1, it can be seen that the salient objects are the mountain tents, which pop out from the background.

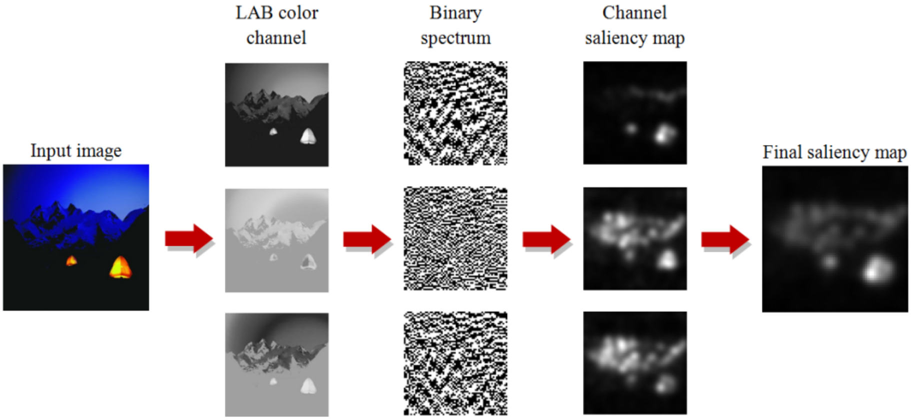

Fig. 1. An illustration of the BWHT method from input image to final saliency map

3 Experimental Validation

In this section, we present the experiments and quantify the consistency of our saliency method with eye fixation data. We compare our method to six popular state-of-the-art saliency approaches in literature by providing an objective evaluation as well as the visual comparison of all saliency maps.

To validate the saliency maps generated by our method, we use the data set of 120 color images from an urban environment and corresponding human eye-fixation data from 20 subjects provided by Bruce and Tsotsos [3]. These color images consist of indoor and outdoor scenes, of which some have very salient items, and others have no particular regions of interest. In order to quantify the consistency of a particular saliency map with a set of fixations of the image, we employ an objective evaluation metric that is referred to as receiver operating characteristic (ROC) area under the curve (AUC). Note that a number of published papers employed ROC-AUC score to evaluate a saliency map's ability to predict human eye fixations.

Following Tatler et al.'s approach [17], we compute the ROC-AUC score conforming to the following procedure. For one image, the positive point set is composed of the fixated locations from all subjects on that image, whereas the negative point set is composed of the non-fixated locations of the image. Each saliency map is binarized by a particular threshold and thereby considered as a binary classifier. At a particular threshold level, a binary saliency map can be divided into the target (white) region and the background (black) region. The true positive rate (TPR) is the proportion of the positive points that fall in the target region of the binary saliency map. The false positive rate (FPR) can be calculated in the same way by using the negative point set. Varying the threshold yields an ROC curve of TPRs versus FPRs, of which the area beneath provides a good measure of the capability of the saliency map to accurately predict where human eye fixations occurred on an image. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0. Chance level is 0.5, and perfect prediction is 1.0.

We compare our saliency maps generated from the proposed method to the following published saliency approaches: the original Itti et al.'s saliency model (ITTI) [2], Harel et al.'s graph-based visual saliency (GBVS) [4], Gao et al.'s discriminant center-surround model (DISC), Bruce and Tsotsos's attention model based on information maximization (AIM) [3], Guo and Zhang's phase spectrum of quaternion Fourier transform (PQFT) [7], and Yu et al.'s saliency approach based on pulsed principal component analysis (PPCA) [9]. All of the saliency approaches are based on the original Matlab implementations available on the author's websites.

An important note about these experiments is that the ROC-AUC score is sensitive to the number of fixations we use in calculation. Former fixations are more likely to be driven by bottom-up manner, whereas later fixations are more likely to be influenced by top-down cues [17]. We calculate the ROC-AUC scores for each image with respect to all fixations, and repeat the process but use only the first two fixation points. Table 1 lists the ROC-AUC score averaged over all 120 images for each saliency method. As expected, the ROC-AUC scores with only the first two fixations are higher than those with all fixations. It can be seen that in both tests our BWHT method has the best capability for predicting eye fixations.

Table 1. The ROC-AUC performance of all seven methods

Method	BWHT	PPCA	PQFT	AIM	DISC	GBVS	ITTI
All fixations	**0.7792**	0.7766	0.7751	0.7706	0.7605	0.7127	0.7062
First 2 fixations	**0.7983**	0.7907	0.7846	0.7777	0.7683	0.7267	0.7182

Fig. 2 gives the saliency maps for 6 sample images from the image data set, which provides a qualitative comparison of all saliency methods. A fixation density map, generated for each image by convolution of the fixation map for all subjects with a Gaussian filter, serves as ground truth. Analysing the qualitative results, we can see that BWHT shows more resemblance to the ground truth. The regions highlighted by our proposed saliency method overlap to a surprisingly large extent with those image regions looked at by humans in free viewing. In addition, high contrast straight edges are suppressed to a much great extent using frequency domain approaches. Good performance with respect to color pop-out is also observed with BWHT compared to the other approaches.

We also record the computational time cost per image in a standard desktop computing environment. Table 2 shows each method's Matlab runtime measurements averaged over the data set. It can be noticed that, not only is BWHT the most predictive of fixations, it also runs faster than all competitors in our tests of computational performance. Note that three frequency domain methods (i.e., BWHT, PPCA and PQFT) are significantly faster than others. This is due to their small number of channels and calculations compared to other saliency methods. PPCA employs the PCA transform and has a computational complexity of $O(N^2)$, where N denotes the total number of pixels of the image. PQFT uses the fast Fourier transform has a computational complexity of $O(N\log N)$. Compared to PPCA and PQFT, the computation of BWHT is mainly comprised of the Walsh-Hadamard transform that can essentially be computed using additions and subtractions only. In computational mathematics, the fast Walsh-Hadamard transform requires only $N\log N$ additions or subtractions and thereby its hardware implementation can be much simpler. Compared to the BWHT, which uses only three color channels at a single spatial scale, ITTI and GBVS rely on seven feature channels and multiple spatial scales; AIM uses 25 filters of 1,323 dimensions. Although these approaches can be accelerated with efficient C implementations, the computational complexity of the BWHT is lower, as suggested by the Matlab runtimes. All seven saliency approaches are implemented in the Matlab R2012a environment on such a computer platform as Intel 3.3 GHz CPU with 8 GB of memory.

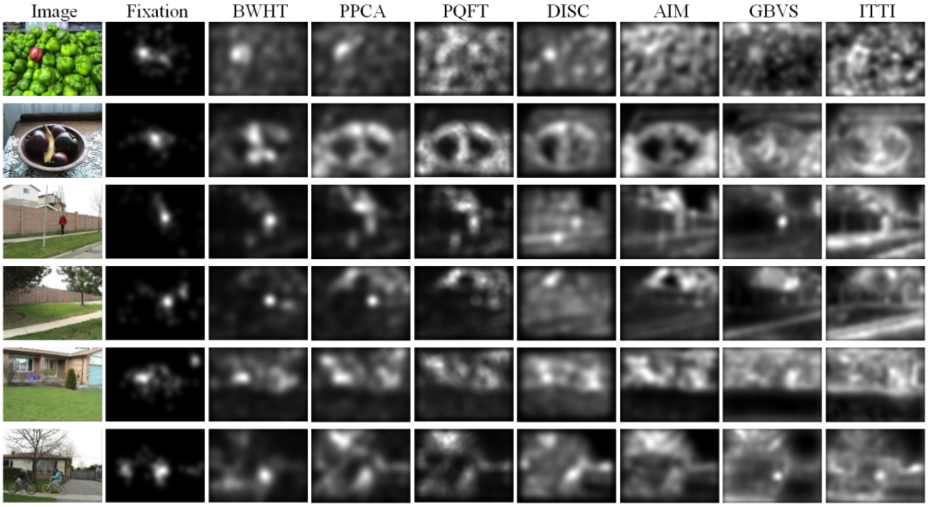

Fig. 2. Qualitative analysis of results for the Bruce data set

Table 2. Computational time cost per image for all seven methods

Method	BWHT	PPCA	PQFT	AIM	DISC	GBVS	ITTI
Time (s)	**0.0018**	0.2337	0.0151	5.0766	1.3778	2.5957	1.1842

4 Conclusions

This paper aims to find a bottom-up visual saliency method based on the Walsh-Hadamard transform. We manifested that the saliency information of an image consists in the binary spectrum of Walsh-Hadamard transform, i.e., the signs of the transform domain coefficients. Experiments in this paper showed that the proposed method is simple and efficient in saliency detection, and outperforms existing state-of-the-art saliency detection approaches. The potentials of our method lies in real-time and interdisciplinary applications focused on computer vision in relation to psychology, robotics and neuroscience.

Acknowledgments. This research was supported by the National Natural Science Foundation of China (Grant No. 61263048), by the Scientific Research Foundation of Yunnan Provincial Department of Education (2012Y277), by the Scientific Research Project of Yunnan University (2011YB21), and by the Young and Middle-Aged Backbone Teachers' Cultivation Plan of Yunnan University (XT412003).

References

1. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Rev. Neurosci. 2(3), 194–203 (2001)
2. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Patt. Anal. and Mach. Intell. 20(11), 1254–1259 (1998)
3. Bruce, N.D., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9(3), 1–24 (2009)
4. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proc. NIPS (2006)
5. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. Journal of Vision 8(7), 1–18 (2008)
6. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Proc. CVPR (2007)
7. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process 19(1), 185–198 (2010)
8. Yu, Y., Wang, B., Zhang, L.: Hebbian-based neural networks for bottom-up visual attention systems. In: Proc. ICONIP (2009)
9. Yu, Y., Wang, B., Zhang, L.: Hebbian-based neural networks for bottom-up visual attention and its applications to ship detection in SAR images. Neurocomputing 74(11), 2008–2017 (2011)
10. Ahmed, N., Rao, K.R.: Walsh-Hadamard transform. In: Orthogonal Transforms for Digital Signal Processing, pp. 99–152. Springer, Heidelberg (1975)

11. Kunz, H.O.: On the equivalence between one-dimensional discrete Walsh-Hadamard and multidimensional discrete Fourier transforms. IEEE Trans.Comput. C-28(3), 267–268 (1979)
12. Li, Z.: A saliency map in primary visual cortex. Trends Cognit. Sci. 6(1), 9–16 (2002)
13. Li, Z., Dayan, P.: Pre-attentive visual selection. Neural Network 19(9), 1437–1439 (2006)
14. Itti, L., Baldi, P.: Bayesian surprise attracts human attention.In: Proc. NIPS (2005)
15. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision 8(7), 1–20 (2008)
16. Barlow, H.B.: Possible principles underlying the transformation of sensory messages. Sensory Communication, 217–234 (1961)
17. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual correlates of fixation selection: effects of scale and time. Vision Research 45(5), 643–659 (2005)