VISUAL SALIENCY BASED ON FAST NONPARAMETRIC MULTIDIMENSIONAL ENTROPY ESTIMATION

Anh Cat Le Ngo[1,2] *Guoping Qiu*[1] *Geoff Underwood*[3] *Li-Minn Ang*[2] *Kah Phooi Seng*[2]

[1] School of Computer Science, University of Nottingham, Jubilee Campus, UK
[2] Faculty of Engineering, University of Nottingham, Malaysia Campus, Malaysia
[3] School of Psychology, University of Nottingham, Unipark Campus, UK

ABSTRACT

Bottom-up visual saliency can be computed through information theoretic models but existing methods face significant computational challenges. Whilst nonparametric methods suffer from the curse of dimensionality problem and are computationally expensive, parametric approaches have the difficulty of determining the shape parameters of the distribution models. This paper makes two contributions to information theoretic based visual saliency models. First, we formulate visual saliency as center surround conditional entropy which gives a direct and intuitive interpretation of the center surround mechanism under the information theoretic framework. Second, and more importantly, we introduce a fast nonparametric multidimensional entropy estimation solution to make information theoretic-based saliency models computationally tractable and practicable in realtime applications. We present experimental results on publicly available eye-tracking image databases to demonstrate that the proposed method is competitive to state of the art.

Index Terms— visual saliency, conditional entropy, k-d tree, information theory, multidimensional entropy estimation.

1. INTRODUCTION

In recent years, there has been increasing interest in the application of the visual saliency mechanism to visual signal processing problems. A predominant theory of computational visual saliency is the center-surround mechanism that is ubiquitously found in the early stages of biological vision [1]. Center-surround saliency models in the time domain [2], frequency domain[3, 4], and information domain [5, 6, 7] have been proposed. Radically, Judd *et al.*[8] proposed that saliency maps can be learned directly from training samples by machine learning methods instead of the center-surround mechanism.

Center-surround methods in the information domain are computationally most challenging because of the curse of dimensionality problem. For instance, both the self-information [6] and the mutual information [7] approaches involve estimating probability density functions in very high dimensional spaces with limited samples. A few work-around solutions have been proposed, by projecting the data onto lower dimensional spaces through independent component analysis (ICA) [6], discrete cosine transform (DCT) [9], and Walsh-Hadamard Transform (WHT) [10] or by modeling information quantity as a parametric Generalized Gaussian Distribution (GGD) [7]. However, projection methods are computationally intensive and estimating the shape parameters of GGD is hard as the authors of [7] have already pointed out.

In this paper, the center surround principle of visual saliency is directly formulated as the conditional entropy of the center given its surrounds. A major contribution of this paper is a fast nonparametric multidimensional entropy estimation solution that overcomes the curse of dimensionality problem and computational complexity issue of information domain visual saliency models thus making information-based saliency models computationally tractable and practicable in real-time applications. We present experimental results on two publicly available eye-tracking still image databases [6, 8] to demonstrate the effectiveness of the proposed method and compare it with existing techniques.

2. SALIENCY BASED ON CENTER SURROUND CONDITIONAL ENTROPY

Let $I_c(x,y)$ be an image patch at location (x,y) and $I_{sr}(x,y)$ its surrounding regions. The conditional entropy of the center given its surround can be defined as $\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) = \boldsymbol{H}(I_c(x,y), I_{sr}(x,y)) - \boldsymbol{H}(I_{sr}(x,y))$ or in terms of joint and marginal probabilities

$$\boldsymbol{H} = \sum_{\substack{I_c(x,y)\in \boldsymbol{I} \\ I_{sr}(x,y)\in \boldsymbol{I}}} p(I_c(x,y), I_{sr}(x,y)) log \frac{p(I_{sr}(x,y))}{p(I_c(x,y), I_{sr}(x,y))} \quad (1)$$

where $\boldsymbol{H}$ is short for $\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y))$. The conditional entropy $\boldsymbol{H}(I_c(x,y)|I_{sr}(x,y))$ can be understood in a number of ways. From a coding or information theory's perspective, it will take $\boldsymbol{H}(I_c(x,y), I_{sr}(x,y))$ bits to code the center and its surrounds together, but if we knew the surround $I_{sr}(x,y)$

already, we will have gained $\boldsymbol{H}(I_{sr}(x,y))$ bits of information, and the conditional entropy measures the remaining bits necessary for coding the center. From an uncertainty or informativeness point of view, the conditional entropy measures the remaining uncertainty of the center once its surrounds are known, or the amount of information of the center given the knowledge of its surrounds. We can use the conditional entropy as a measure of saliency, i.e.

$$\boldsymbol{S}(x,y) = \boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) \quad (2)$$

The definition of saliency in equation (2) and (1) is consistent with a number of definitions in the literature including self-information [6], surprise [11] and decision theoretic saliency [7]. The self-information saliency of [6] measures the self-information of $I_c(x,y)$ in the context of its surrounds, $-log\{p(I_c(x,y))\}$. If $I_c(x,y)$ is a common patch within the image, then $p(I_c(x,y))$ is large, $-log\{p(I_c(x,y))\}$ will be small, hence the saliency will be small. $\boldsymbol{S}(x,y)$ in (2) has the same property, that is, if the center and its surrounds are very similar, then $\boldsymbol{S}(x,y)$ will be small and vice versa. The surprise measure of [11] can be re-written as ($\boldsymbol{S}$ is short for $\boldsymbol{S}(I_c(x,y), I_{sr}(x,y))$

$$\boldsymbol{S} = \sum_{\substack{I_c(x,y)\in \boldsymbol{I} \\ I_{sr}(x,y)\in \boldsymbol{I}}} p(I_{sr}(x,y)) log \frac{p(I_{sr}(x,y))}{p(I_{sr}(x,y)|I_c(x,y))} \quad (3)$$

Here, the surrounds $I_{sr}(x,y))$ can be interpreted as the model or background information and the center $I_c(x,y)$ as the new observation data. Again, the surprise measure will be small when the center and surround are similar and large when they are different. The decision theoretic discriminant saliency of [7] boils down to the computation of mutual information between the center and its surround, while mutual information and conditional entropy are related as follows.

$$\boldsymbol{MI}(I_c(x,y), I_{sr}(x,y)) = \boldsymbol{H}(I_c(x,y)) - \boldsymbol{H}(I_c(x,y)|I_{sr}(x,y)) \quad (4)$$

$\boldsymbol{MI}(I_c(x,y), I_{sr}(x,y))$ is deduced amount of uncertainty for the center $I_c(x,y)$ if its surrounds $I_{sr}(x,y)$ are known. MI can be interpreted as how much similarity surround and center data has, therefore it is consistent with conditional entropy. A large mutual information means significant overlap between center and surround information hence the saliency is small, so is the conditional entropy. All these information measurements involve the estimation of probability density functions in very high dimensional spaces with limited data samples, a very challenging problem. In practice, various simplification processes have to be used, e.g., [6] employed independent component analysis (ICA) and [7] assumed a parametric Generalized Gaussian Distribution (GGD) model. In the next section, we introduce a fast non-parametric method.

3. FAST NONPARAMETRIC ESTIMATION OF CENTER SURROUND CONDITIONAL ENTROPY

Visual data have excessive amount of information, but only some attracts attention at early stage. Itti *et al.*[2] used low-level features of intensity, colour and orientation at multi-resolution to build several conspicuity maps and combine them linearly to form a saliency map. In the discriminant saliency map approach, Gao *et al.*[7] extracted and modeled band-pass features by Wavelet/Gabor Filters and used parametric GGD to estimate the mutual information between the center and surround. In this paper, we use mid-band frequency features which have been shown to allow the best prediction of attention globally [12]. Figure 1 shows a step-by-step illustration of mid-band filtering. Firstly, a 9/7 Cohen-Daubechies-Feauveau (CDF) wavelet[13] decomposes an image by three levels. Then, all level 1 components and level 3 low-low frequency component are removed. Finally, the remaining components are converted back to time-domain by the inverse of the 9/7 CDF wavelet to form the mid-band image. The mid-band image is divided into NxN patches (8x8 patches are utilized in this paper). The saliency of each center patch **C**, is computed as the conditional entropy of **C** given four of its surrounding patches (**N**, **S**, **W**, and **E**) as

Fig. 1. Medium Band Filter Flow Chart

$$\begin{aligned} S(\mathbf{C}) &= H(\mathbf{C}|(\mathbf{N},\mathbf{S},\mathbf{W},\mathbf{E})) \\ &= H(\mathbf{C},\mathbf{N},\mathbf{S},\mathbf{W},\mathbf{E}) - H(\mathbf{N},\mathbf{S},\mathbf{W},\mathbf{E}) \end{aligned} \quad (5)$$

Estimating the two joint entropies on the right-hand side of (5) is challenging because of the high dimensionality of the data. To get round the problem, we take a similar approach as [14] and treat the coordinate locations of the pixels as random variables and approximate (5) as

$$\begin{aligned} S(\mathbf{C}) \quad &= \quad H(c(x,y), n(x,y), s(x,y), w(x,y), e(x,y)) \\ &- \quad H(n(x,y), s(x,y), w(x,y), e(x,y)) \end{aligned} \quad (6)$$

where $c(x,y), n(x,y), s(x,y), w(x,y), e(x,y)$ are respectively pixels from the $\mathbf{C}, \mathbf{N}, \mathbf{S}, \mathbf{W}, \mathbf{E}$ patches at the same reference location (*x*,*y*). We treat the problem as drawing samples from (*x*,*y*) in order to approximate the conditional entropy. With the formulation of (6), we can now simplify the problem as estimating the entropies in the 4-D and 5-D spaces with a total of 8x8 = 64 samples. We use a technique similar to [15] to achieve fast implementation of (6). The technique is based on a k-d tree style approach to partitioning the input data space $\Omega \in \Re^D$ into $A = \{A_j | j = 1, 2, \ldots, m\}$ with $A_i \bigcap A_j = \varnothing$ if $i \neq j$ and $\bigcup_j A_j = \Omega$. Let n_j be the number of samples in the cell A_j, $V(A_j)$ the volume of cell A_j, the total number of samples *N*, then the multidimensional joint entropy can be estimated as

$$\hat{H} = \sum_{j=1}^{m} \frac{n_j}{N} log\left(\frac{N}{n_j} V(A_j)\right) \quad (7)$$

Fig. 2. From left to right: Image Sample, ITT, AIM, ENT and MIT saliency maps

The computational complexity of the algorithm is $\Theta\left(DNlogN\right)$ and the space complexity is $\Theta(DN)$. For the algorithm to work, the sample size has to satisfy $N \geq 2^D$. Our setting, $N = 64$ and $D = 5$ or $D = 4$, $2^5 = 32$ and $2^4 = 16$, therefore meets the samples size requirement of the algorithm.

4. EXPERIMENTAL RESULTS

We evaluate the new conditional entropy based saliency method (from now on referred to as ENT) on publicly available eye tracking databases of Bruce and Tsotos [6] and Judd *et al.*[8], and compare it with a number of saliency estimation methods in the literature including, Itti and Koch (ITT) [2], spectral residual saliency (SRS) [3], Information Maximization (AIM)[6], and discriminant saliency (DIS) [7]. Fig. 2 shows the saliency maps generated by different methods of a sample image. It is seen that the visual appearance of these saliency maps are quite similar. To compare the performances of different methods quantitatively, we use Tatler's numeric measurement [16]. Saliency maps are treated as binary classifiers to discriminate fixation points versus non-fixation points. Threshold for classifying fixation points are not fixed but systematically changed from minimum to maximum of saliency values to generate ROC curves. ROC curves of various methods on Bruce's database [17] are shown in Fig. 3. Area under the ROC curves (AUC) have been used by a number of authors to give quantitative comparison of saliency computation methods and table 1 shows the AUC values of five different methods.

Table 1. Area Under Curve (AUC) for different methods.

Methods	ITT [2]	AIM [6]	New ENT	DIS[7]	SRS[3]
AUC	0.70947	0.73873	0.78167	0.76940	0.75434

The ROC curves show that the new ENT method generally performs better than AIM and ITT methods, and the performances are reconfirmed by the area under curve (AUC) results in table 1. In the table, the AUC result of DIS saliency method was performed on the same database by the original authors and taken directly from [7]. These AUC results show that ENT methods also performs better or at least as well as the DIS method.

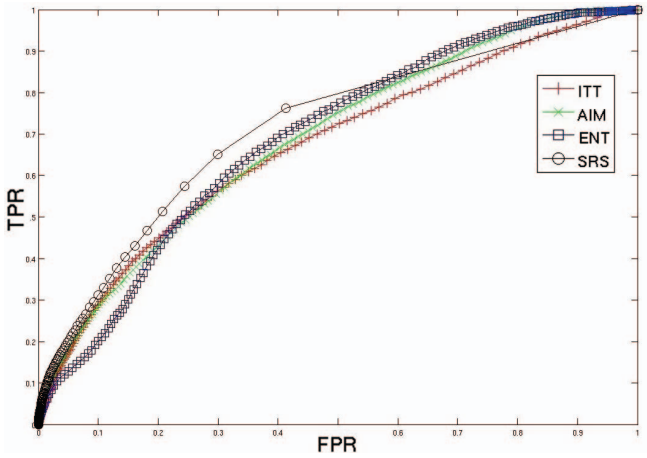

Fig. 3. ROC of ITT, AIM, ENT and SRS methods

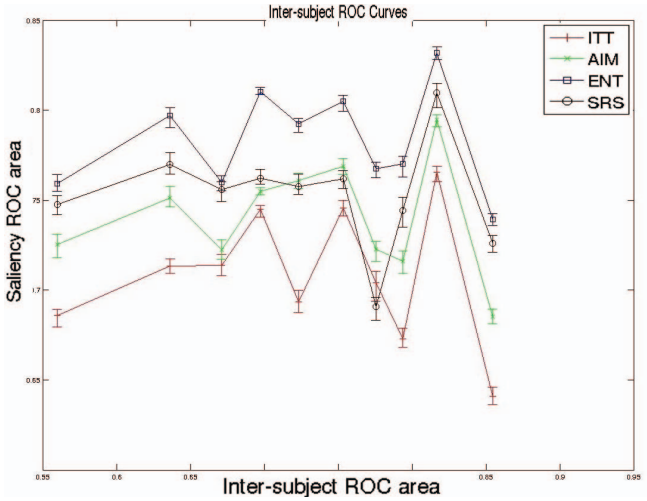

Fig. 4. Inter-subject ROC of ITT, AIM, ENT, and SRS

ROC curves and AUC values are useful for comparing different computational saliency approaches, but they do not show relationships between these methods and eye tracking data. Inter-subject ROC curves proposed by Harel *et al.*[18] help to show performances of human visual system versus that of computational saliency methods. Fig. 4 shows the Inter-subject ROC curves of ENT against a few other methods for the database of [6]. This plot clearly demonstrates that our new ENT technique has outperformed current state-of-art saliency methods and displayed good matching with eye-tracking data.

Table 2. Time Consumption of Saliency Methods

Methods	ITT	AIM	ENT	SRS
Time (s)	1.2488	66.2673	0.93094	0.33654

Information-theoretic saliency methods such as AIM has drawbacks due to their intensive computational requirements and are unsuitable for real-time applications. The proposed method can overcome this issue. Table 2 shows the computational speeds of several techniques. It is seen that ENT is

over 70 times faster than AIM and 1.3 times faster than ITT. Though it is slower than SRS, ENT can better match eye-fixation data than SRS. All experiments are done in MATLAB on a 2.33 GHz Intel Core 2 Duo computer running Linux.

The ENT method was further tested on the database created by Judd *et al.*[8]. Quantitative results in table 3 once again show that our method compares very well against other state of the art methods.

Table 3. Area Under Curve (AUC)

Methods	ITT	AIM	ENT	MIT[8]
AUC	0.74940	0.71165	0.78157	0.68845

5. CONCLUDING REMARKS

In this paper, we have formulated center-surround bottom-up visual saliency as conditional entropy and presented a fast nonparametric multidimensional entropy estimation solution to overcome the inherent curse of dimensionality in conventional nonparametric approaches and difficulty of determining shapes of distribution in parametric approaches of information domain visual saliency models. We have shown that the new method is not only computationally efficient but also achieves state of the art performances on publicly available eye tracking databases.

6. REFERENCES

[1] D.H. Hubel and T.N. Wiesel, "Receptive fields and functional architecture into nonstriate visual areas (18 and 19) of the cat.," *Journal of neurophysiology*, vol. 28, pp. 229–89, Mar. 1965.

[2] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*. 2007, number 800, pp. 1–8, Citeseer.

[4] C.L. Guo, Q. Ma, and L.M. Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[5] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[6] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, pp. 155, 2006.

[7] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1–8, 2007.

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

[9] G. Qiu, X. Gu, Z. Chen, Q. Chen, and C. Wang, "An information theoretic model of spatiotemporal visual saliency.," in *Proc. of ICME2007*, 2007, pp. 1806–1809.

[10] L. Zheng, G. Qiu, J. Huang, and H. Fu, "Salient covariance for near-duplicate image and video detection," in *Proc. of ICIP2011*, 2011, pp. 2585–2588.

[11] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Advances in neural information processing systems*, vol. 18, pp. 547, 2006.

[12] F. Urban, B. Follet, C. Chamaret, O. Meur, and T. Baccino, "Medium Spatial Frequencies, a Strong Predictor of Salience," *Cognitive Computation*, Nov. 2010.

[13] A. Cohen, I. Daubechies, and J.C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 45, no. 5, pp. 485–560, June 1992.

[14] P. Viola and W.M. Wells III, "Alignment by maximization of mutual information," in *ICCV*. 1995, vol. 24, p. 16, Published by the IEEE Computer Society.

[15] D. Stowell and M.D. Plumbley, "Fast Multidimensional Entropy Estimation by k-d Partitioning," *Signal Processing Letters, IEEE*, vol. 16, no. 6, pp. 537–540, 2009.

[16] B.W. Tatler, R.J. Baddeley, and I.D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time.," *Vision research*, vol. 45, no. 5, pp. 643–59, Mar. 2005.

[17] N. Bruce and J.K. Tsotsos, "Saliency , attention , and visual search : An information theoretic approach," *Journal of Vision*, vol. 9, pp. 1–24, 2009.

[18] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, pp. 545, 2007.