Examining Visual Saliency Prediction in Naturalistic Scenes

Conference Paper · October 2014

DOI: 10.1109/ICIP.2014.7025829

4 authors, including:

Shafin Rahman
North South University
13 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Neil D. B. Bruce
University of Manitoba
46 PUBLICATIONS **1,103** CITATIONS
SEE PROFILE

EXAMINING VISUAL SALIENCY PREDICTION IN NATURALISTIC SCENES

Shafin Rahman, Mrigank Rochan, Yang Wang, Neil D.B. Bruce *

Department of Computer Science
University of Manitoba, Canada
{shafin12,mrochan,ywang,bruce}@cs.umanitoba.ca

ABSTRACT

Given the significant number of potential applications, visual saliency has increasingly become an area of interest in image and vision research. Many different strategies for predicting visual saliency have been proposed, that differ in their composition or rationale, and with a significant focus on improving performance across standard benchmarks. Recent benchmarks considering a large number of algorithms have further provided an understanding of the behavior of different algorithms. Performance evaluation has primarily focused on indoor and outdoor images of urban environments, many of which are composed, and contain salient objects. In this work, we test the performance of a number of the better performing algorithms on data derived from naturalistic scenes. In addition, given the strong connection to human vision, we test a putative model for early visual processing in primates tied to spectral energy and normalization. Results demonstrate significant differences between common datasets, and natural images. Performance analysis of the second-order contrast model also provides additional insight concerning the role of spectral energy in determining saliency. Finally we include analysis that demonstrates statistical properties of images that tend to imply common gaze patterns across observers.

Index Terms— visual saliency, fixation, natural image statistics, benchmarking, spectral energy

1. INTRODUCTION

Humans demonstrate a great capacity to locate the region of interest within a very short time, and to focus processing on relevant stimuli or patterns. This capability is important to active vision systems in artificial vision, and increasingly important in computational application domains including proto-object detection, tracking, recognition, video summarization, image compression, navigation, and advertising. In addressing this need, a significant number of proposals to characterize visual saliency have emerged over the past decade.

Models of attention are typically classified as either top-down or bottom up [1, 2] with the distinction of being task-driven or stimulus-driven respectively. Attention involves a complex interplay of both task and stimulus driven forces, which presents a challenging model problem in it's most general form. Many studies on attention have focused on tracking gaze in *free viewing* paradigms, wherein a user is presented with a series of images without a particular task instruction. This carries the assumption that gaze patterns will tend to be more strongly driven by stimulus patterns. Visual saliency models have largely been tested on gaze data from this type of experiment, but images appearing in most data sets tend to have significant context, objects or salient regions that may imply the involvement of higher-level cognitive routines. For this reason, it is important to understand the behavior of models within more naturalistic settings wherein there are weaker priors on context, or objects present. To this end, in this paper, we evaluate a number of algorithms that have performed well in recent benchmarking studies, subjected to natural images.

For such naturalistic stimuli, it is also natural to consider computational saliency from the perspective of visual information processing in human brain. To this end, we also test a recent simplistic biologically motivated model that is supported by it's agreement with brain imaging data [3]. A simplistic bio-inspired model of early neural activation in the visual cortex is based on log-Gabor filters with activation corresponding to the sum of spectral energy represented across filters corresponding to a particular location [3]. This model is motivated by the complex cell energy model wherein energy corresponds to paired oriented Gabor-like filters in quadrature. This simplistic model is successful in predicting blood oxygenation level dependent (BOLD) responses in viewing natural images and checkerboard patterns, however responses get saturated for low contrast energy. An improvement to this approach is proposed by Kay *et al.* [4] wherein divisive normalization (DN) and second order contrast (SOC) follow the computation of total spectral energy to achieve better prediction of BOLD responses. Therefore, it is of interest to examine the extent to which energy filters coupled with DN and SOC provides a characterization of fixated locations for naturalistic stimuli. This model has therefore also been implemented and included in our comparison.

Substantial research has been conducted on modeling natural image structure by means of statistical analysis [5] but the

*The authors gratefully acknowledge financial support from NSERC Canada and the University of Manitoba.

connection of natural images to saliency and fixation patterns has not been addressed in recent benchmarks. We therefore provide a benchmark for naturalistic images, including in addition a test of the efficacy of the SOC model in predicting fixation data, and examine how statistical properties of images relate to expected viewing patterns.

2. BACKGROUND

To provide a comparison of algorithms for naturalistic stimuli, we have employed the DOVES dataset containing only grayscale natural images with no salient objects and also having fixation data from a large number of observers [6]. This dataset contains 101 calibrated grayscale natural images of resolution 1024×768 and has 30,000 fixation points from 29 participants. Very few algorithms in the literature use this dataset for attention modeling [7, 8]. In addition, as our results demonstrate, benchmark scores presented in prior work are misleading given the impact that spatial bias in the data may carry. We have also employed the dataset of Bruce and Tsotsos [9] to further examine the behaviour of the SOC model for standard images. This dataset has 120 color indoor and outdoor images and eye tracking data from 20 participants.

Algorithms selected for comparison are based on a representative sample of available algorithms that demonstrate strong performance in the benchmark of Borji *et al.* [1]. While complete coverage of the details of these algorithms is not possible in this format, the reader is urged to consult [1] for a summary, or the references listed in Table 1 for original sources.

2.1. Spectral Energy and Second Order Contrast

Models of early visual processing in primates has been implicated as having an important role in directing visual attention [10]. In line with developing a deeper understanding of early visual processing, a great deal of effort has been directed at testing putative models for predicting the responses of neurons [3] in the primary visual cortex (V1). One approach to matching models to behavioral observations is had in examining blood oxygenation level dependent (BOLD) responses in localized regions (voxels) of the visual cortex. Recent work has presented a model that involves a number of components consisting of localized oriented band-pass filtering (via Gabor filters), divisive normalization, and second order contrast (SOC). Given the relevance of this work to visual saliency, we have also included an implemented model that draws inspiration from neurobiology [4] for comparison with a number of the better performing models across standard benchmarks. Details of the implemented SOC model are as follows:

$$F_{v1}(x,y) = exp\frac{-\log(x'/y')^2}{2\log(\sigma_x/x')^2}.exp\frac{-y'^2}{2\sigma_y^2} \quad (1)$$

where, $x' = x\cos(\theta) + y\sin(\theta), y' = -x\sin(\theta) + y\cos(\theta)$, θ is the orientation of the filter, x' is the center frequency and σ_x and σ_x are the bandwidth of the filter along the x and y direction respectively. Using different scale and orientation a bank of filter can be produced. Subsequent summation of filter responses provides a basic saliency map.

We can improve on this model using some steps from a recent study presented in [4]. The prediction of BOLD responses using Gabor like filters can not model the fact that responses saturate at low contrast. To solve this problem, Kay *et al.* [4] proposed divisive normalization (DN) as an additional component of importance. This step is able to capture several nonlinear response properties of V1 neurons. Computationally, this step can be described by the following equation:

$$D_n = \frac{G^r_{sc,or}}{s^r + (\frac{\Sigma_{sc,or}G_{sc,or}}{S\times O})^r} \quad (2)$$

where, $G_{sc,or}$ is the log-Gabor filter response for scale sc and orientation or; S and O are the total number of scales and orientations respectively; s and r control the strength of normalization.

As noted [4], this model is still incomplete because it overestimates actual BOLD responses. Computing the spatial variance of contrast energy provides predictions in line with BOLD responses to visual patterns. This is called second order contrast (SOC) step. SOC is included in the computation performed as:

$$SOC = \Sigma_i w_i(D_n - c\Sigma_j w_j D_j) \quad (3)$$

where, w is 2D Gaussian weights, D_j is the Gaussian filtered image of D_i and c controls the strength of non-linearity of second order contrast. SOC contains the final saliency map.

3. EXPERIMENTAL METHODS

As stated, a central goal of the current work is to examine the extent to which algorithms that compute visual saliency are predictive of viewing patterns in natural scenes. We have therefore compared the saliency output of various algorithms as compared with fixation eye tracking data via the standard of using ROC curves for performance benchmarking. This evaluation includes 10 of the best performing algorithms from the study of Borji et al. [1] as well as the SOC model with performance based on average area under the ROC curve (auROC) across all images in the dataset. There is some contention as to the best method for evaluation, and as such we have employed two different types of implementation of this type of evaluation. The first follows the evaluation process of Judd *et al.* [2] and the second using the shuffled AUC process employed in [11, 1]. In the latter case, negative samples are randomly sampled from fixated locations of images other than that being evaluated to remove the impact of center bias.

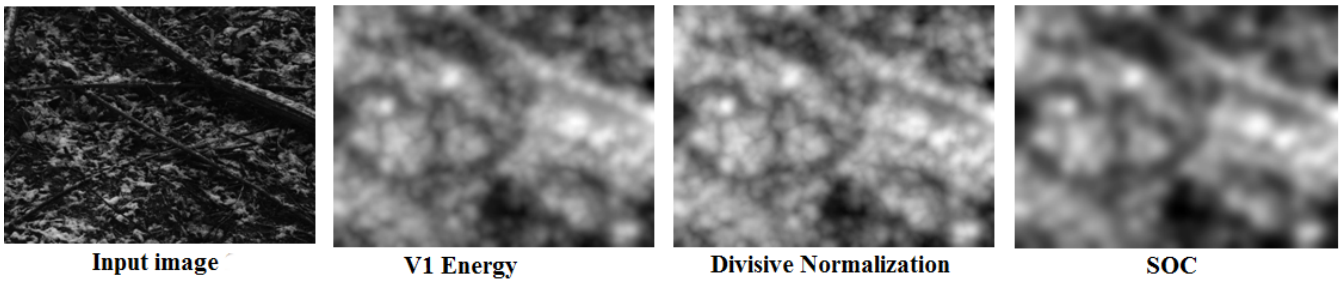

Fig. 1. Various stages of the SOC processing cascade for a naturalistic image.

4. RESULTS

The results presented serve to provide a sense of performance of various saliency models in predicting fixations for naturalistic images, in addition to validating the SOC model on the basis of it's ability to predict the locus of fixated locations.

In the context of the SOC model spectral energy is characterized by log-Gabor filters in quadrature, and subject to four parameters (minimum wavelength, a multiplier that defines the centre frequency for any given scale, the σ of the Gaussian envelope containing the log-Gabor filer and total number of scales). The total number of orientation was fixed to 6 as behavior of the algorithm was found to be relatively insensitive to this parameter. Parameters for energy filtering were selected from a broad sample of parameter combinations to produce a best fit to the characterization of fixation locations on the DOVES dataset. In particular, minimum wavelength was chosen from 2, 3, 6, 12 pixels, the multiplier associated with peak frequency of bandpass filters from one scale to the next 1.4, 1.7, 2.2, the sigma on the Gaussian envelope (normalized by frequency) from .75, .65, .55 and number of scales from 3,4.The best parameter set was found to be 12, 1.7, 0.65, 4 for the four respective parameters. It is important to note that this provides even spectral coverage over a broad band of radial frequencies.

Subsequently, parameters for divisive normalization (as defined in the SOC model) involved optimizing r and s in choosing the best r from .7, .9, 1, 1.2 and s from .5, 4, 8, 16. The best results were obtained in choosing r = .7 and s = 8. Finally second order contrast (SOC) parameters were optimized with the 2 parameters σ and c optimized selecting the best c among 0.5, 0.7, 0.9 and σ among 2, 4, 8, 16. Optimal results were obtained in choosing σ = 2, c = 0.5. As an additional test of fidelity (and to avoid the possibility that parameters are overfit to a specific dataset, we have also compared the results against the best performing algorithms in prior benchmarks on the Bruce and Tsotsos dataset [9, 12] with the same parameters optimized for the DOVES dataset.

A comparison of performance of algorithms that fare well in the benchmark of Borji et al. in addition to the SOC model is presented in Table 4. We have use two different dataset for this comparison, specifically the DOVES and Bruce *et al.* dataset, with the latter serving as an additional test of the simplistic SOC model. It is evident that the performance for the DOVES dataset is significantly lower than that of the Bruce *et al.* dataset. This most likely reflects the observation that the DOVES dataset is relatively devoid of very salient objects. It is also the case that the performance of the relatively simplistic SOC model is within a similar range to some of the better performing saliency algorithms, despite it's simplicity. It is important to also note that the SOC model operates only on a grayscale version of the Bruce and Tsotsos dataset, and is therefore at a disadvantage in information provided as input but nevertheless performs in a range that is proximal to the state-of-the-art.

It is important to note that the large differences in the Judd AUC scores may be attributed primarily to the extent to which there is central bias in the algorithms evaluated. Based on the shuffled AUC score, we have computed the inter-observer average auROC for the DOVES dataset to be 0.5526 ± 0.04518 with a maximum of 0.6418, and for the Bruce and Tsotsos dataset to be 0.7372 ± 0.0918, with a maximum of 0.8768. This suggests that the saliency algorithms tested are within the range of the inter-observer prediction scores for the naturalistic data, but fall short of the inter-observer predictions for the Bruce and Tsotsos data. Inter-observer predictions were computed on a leave-one-out basis for all observer/image combinations, with fixations from remaining observers convolved with a Gaussian to produce a saliency map. The optimal σ for the Gaussian convolution was determined based on the average auROC across all individual/image combinations to provide with the best σ on average used to compute IO scores. This corresponded to a σ of 3% of the image width.

4.1. Saliency and Inter-Observer Agreement

It is evident that the performance for the naturalistic stimuli present in the DOVES dataset is significantly lower than alternative datasets. This may reflect the relative absence of very salient items within this dataset. With that said, there is significant variability in performance, and some individual images have quite respectable auROC scores across the range of algorithms. It is natural to consider some simple quantifiable measures that may be correlated with image content, and

Algorithms [1]	DOVES		Bruce *et al.*	
	Judd AUC	Shuffled AUC	Judd AUC	Shuffled AUC
Torralba [13]	.582	.556	.825	.673
HouCVPR [14]	.557	.551	.780	.675
HouNIPS [15]	.655	.539	.780	.661
Yin Li [16]	.594	.525	.698	.556
Itti-CIO2 [17]	.808	.566	.808	.647
ImageSignatureLab [18]	.612	.550	.815	.698
ImageSignatureRGB [18]	.612	.549	.796	.677
SDSR [19]	.543	.553	.814	.707
AIM [9, 12]	.562	.559	.827	.685
GBVS [20]	.777	.549	.825	.615
Yan [21]	.691	.553	.810	.671
AWS [22]	.651	.538	.808	.711
SOC (inspired by [4])	.573	.556	.789	.652

Table 1. Comparison of algorithms for DOVES [6] and Bruce [9] datasets with a representative sample of algorithms that perform well across other benchmarks [1].

Fig. 2. Images showing the highest inter-observer auROC scores among naturalistic images.

also with the expectation of more commonality in the focus of attention. As many of the more successful models appeal to information, coding efficiency or compression of patterns within a scene, one natural avenue to consider is statistics tied to the aforementioned factors.

In this analysis, we draw inspiration from the Image Signature model [18], which makes its determination based on casting the image into a representation based on DCT coefficients. We have examined various statistics of DCT coefficients with respect to their diagnosticity in predicting images that are likely to have common focal points among human observers. A summary of statistics computed on the DCT coefficients, and corresponding Pearson's ρ with corresponding significance values p measured against per image inter-observer ROC scores appears in table 4.1.

Statistic	**Pearson's ρ**	**p-value**
mean(c_k)	0.1224	0.2227
std(c_k)	0.1996	0.0454
skew(c_k)	0.2886	0.0034
kurtosis(c_k)	0.2876	0.0035
$l^1(c_k)$	-0.1775	0.0792
$l^2(c_k)$	-0.1519	0.1295
$l^0_\epsilon(c_k), \epsilon = 5$	0.2214	0.0261
$\sum -log(1 + c_k^2)$	0.2014	0.0434
$\sum c_k^4/(\sum c_k^2)^2$	0.2876	0.0035

Table 2. Correlation of statistics of DCT coefficients c_k with inter-observer auROC

5. CONCLUSION

In this paper, we have examined the performance of various algorithms that perform well across standard saliency benchmarks for naturalistic stimuli. We have also tested a specific bio-inspired hypothesis for early visual computation as an additional comparator.

Results indicate that models do not perform as well for naturalistic images where there may be less strong salient targets. Results suggest that similar performance to some of the better algorithms for visual saliency computation can be obtained via a relatively simplistic model. We have also demonstrated that basic statistical properties of an image provide a reasonable basis for prediction of commonality in viewing patterns across human observers.

6. REFERENCES

[1] A. Borji, D.N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[2] Tilke Judd, Fredo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," *MIT Computer Science and Artificial Intelligence Lab Technical Report*, 2012.

[3] A. Garcia-Diaz, X.R. Fdez-Vidal, X.M. Pardo, and R. Dosil, "Decorrelation and distinctiveness provide with human-like saliency," *Lecture Notes in Computer Science*, vol. 5807 LNCS, pp. 343–354, 2009.

[4] K.N. Kay, J. Winawer, A. Rokem, A. Mezer, and B.A. Wandell, "A two-stage cascade model of bold responses in human visual cortex," *PLoS Computational Biology*, vol. 9, no. 5, 2013.

[5] D.J. Field, "Relations between the statistics of natural images and the response properties of cortical cells.," *Journal of the Optical Society of America. A, Optics and image science*, vol. 4, no. 12, pp. 2379–2394, 1987.

[6] I. Van Der Linde, U. Rajashekar, A.C. Bovik, and L.K. Cormack, "Doves: A database of visual eye movements," *Spatial Vision*, vol. 22, no. 2, pp. 161–177, 2009.

[7] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 473–480, 2011.

[8] J. Liu and Y. Liu, "A model for saliency detection using nmfsc algorithm," *Lecture Notes in Computer Science*, vol. 5702 LNCS, pp. 301–308, 2009.

[9] Neil D. B. Bruce and John K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[10] Zhaoping Li, "A saliency map in primary visual cortex," *Trends in cognitive sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[11] 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, no. 7, 2008.

[12] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[13] Antonio Torralba, "Modeling global scene factors in attention," *JOSA A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[14] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[15] Xiaodi Hou and Liqing Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 2008, pp. 681–688.

[16] Yin Li, Yue Zhou, Junchi Yan, Zhibin Niu, and Jie Yang, "Visual saliency based on conditional entropy," in *Computer Vision–ACCV 2009*, pp. 246–257. Springer, 2010.

[17] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] Xiaodi Hou, Jonathan Harel, and Christof Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 194–201, 2012.

[19] Hae Jong Seo and Peyman Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, vol. 9, no. 12, 2009.

[20] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[21] Junchi Yan, Jian Liu, Yin Li, Zhibin Niu, and Yuncai Liu, "Visual saliency detection via rank-sparsity decomposition," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1089–1092.

[22] Anton Garcia-Diaz, Xose R. Fdez-Vidal, Xose M. Pardo, and Raquel Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51 – 64, 2012.