Saliency Based on Decorrelation and Distinctiveness of Local Responses

Antón Garcia-Diaz, Xosé R. Fdez-Vidal, Xosé M. Pardo, and Raquel Dosil

Universidade de Santiago de Compostela, Grupo de Visión Artificial,
Departamento de Electrónica e Computación, Campus Sur s/n,
15782 Santiago de Compostela, Spain
{anton.garcia,xose.vidal,xose.pardo,raquel.dosil}@usc.es

Abstract. In this paper we validate a new model of bottom-up saliency based in the decorrelation and the distinctiveness of local responses. The model is simple and light, and is based on biologically plausible mechanisms. Decorrelation is achieved by applying principal components analysis over a set of multiscale low level features. Distinctiveness is measured using the Hotelling's T^2 statistic. The presented approach provides a suitable framework for the incorporation of top-down processes like contextual priors, but also learning and recognition. We show its capability of reproducing human fixations on an open access image dataset and we compare it with other recently proposed models of the state of the art.

Keywords: saliency, bottom-up, attention, eye-fixations.

1 Introduction

It is well known, from the analysis of visual search problems, that vision processes have to face a huge computational complexity [1] . The Human Visual System (HVS) tackles this challenge through the selection of information with several mechanisms, starting from foveation. In the basis of this selection is the visual attention, including its data-driven component leading to the so called bottom-up saliency. In the last decades, the interest in the understanding of saliency mechanisms and the appraisal of its relative importance in relation to the top-down (knowledge-based) relevance, has constantly raised. Besides, attention models can facilitate a solution of technical problems, ranging from robotics navigation [2] to image compression or object recognition [3].

Recently, several approaches to bottom-up saliency have been proposed based on similarity and local information measures. In these models local distinctiveness is obtained either from self-information [4][5], mutual information [6][7], or from dissimilarity [8], using different decomposition and competition schemes.

In this paper, we propose a new model of bottom-up saliency, simple and with low computational complexity. To achieve this, we take into account the decorrelation of neural responses when considering the behavior of a population of neurons subject to stimuli of a natural image [9]. This is believed to be closely related to the important role of non classical receptive fields (NCRF)

X. Jiang and N. Petkov (Eds.): CAIP 2009, LNCS 5702, pp. 261–268, 2009.

in the functioning of HVS. Therefore, we start from a multiscale decomposition on two feature dimensions: local orientation energy and color. We obtain the decorrelated responses applying PCA to the multiscale features. Then, we measure the statistical distance of each feature to the center of the distribution as the Hotelling's T^2 distance. Finally, we apply normalization and Gaussian smoothing to gain robustness. The resulting maps are firstly summed, delivering local energy and color conspicuities, and then they are normalized and averaged, producing the final saliency map.

In order to achieve a psychophysical validation, most models of bottom-up saliency assess their performance in predicting human fixations, and compare their results with those provided by other previous models. The most frequent comparison method consists in the use of the receiver-operator-curve (ROC) and the corresponding value of the area under the curve (AUC), as a measure of predictive power [5][6][8]. The use of Kullback-Leibler (K-L) divergence to compare priority and saliency maps is also found on related literature [10]. From the use of both methods, we obtain results that match or improve those achieved with models of the state of the art. Moreover, ruling out the use of center-surround differences we definitely improve the results respect to those obtained with a previous proposal [11].

The paper is developed as follows. Section 2 is devoted to overview the visual attention model. In Section 3 we present and discuss the experimental work carried out, and the achieved results. Finally, Section 4 summarizes the paper.

2 Model

Our model takes as input a color image codified using the Lab color model. Unlike other implementations of saliency [6][12] this election is based on a widely used psychophysical standard. We decompose the luminance image by means of a Gabor-like bank of filters, in agreement with the standard model of V1. Since orientation selectivity is very weakly associated with color selectivity, the opponent color components a and b simply undergo a multiscale decomposition. Hence, we employ two feature dimensions: color and local energy. By decorrelating the multiscale responses, extracting from them a local measure of variability, and further performing a local averaging, we obtain a unified and efficient measure of saliency.

2.1 Local Energy and Color Maps

Local energy is extracted applying a bank of log Gabor filters [13] to the luminance component. In the frequency domain, the log Gabor function takes the expression:

$$logGabor(\rho, \alpha; \rho_i, \alpha_i) = e^{-\frac{(\log(\rho/\rho_i))^2}{2(\log(\sigma_{\rho i}/\rho_i))^2}} e^{-\frac{(\alpha-\alpha_i)^2}{2(\sigma_\alpha)^2}}. \quad (1)$$

where (ρ, α) are polar frequency coordinates and (ρ_i, α_i) is the central frequency of the filter. Log Gabor filters, unlike Gabor, have no DC or negative frequency

components, therefore avoiding artifacts. Their long tail towards the high frequencies improves its localization. In the spatial domain, they are complex valued functions (with no analytical expression), whose components are a pair of filters in phase quadrature, f and h. Thus for each scale s and orientation o, we obtain a complex response. Its modulus is a measure of the local energy of the input associated to the corresponding frequency band [14] [15] .

$$e_{so}(x,y) = \sqrt{(L * f_{so})^2 + (L * h_{so})^2}. \tag{2}$$

Regarding the color dimension, we obtain a multiscale representation both for a and b, from the responses to a bank of log Gaussian filters.

$$logGauss(\rho) = e^{-\frac{(\log(\rho))^2}{2(\log(2^n\sigma))^2}}. \tag{3}$$

Thus, for each scale and color opponent component we get a real valued response. The parameters used here were: 8 scales spaced by one octave, 4 orientations (for local energy), minimum wavelength of 2 pixels, angular standard deviation of $\sigma_\alpha = 37.5^o$, and a frequency bandwidth of 2 octaves.

2.2 Measurement of Distinctiveness

Observations from neurobiology show decorrelation of neural responses, as well as an increased population sparseness in comparison to what can be expected from a standard Gabor-like representation [16]. Accordingly we decorrelate the multiscale information of each sub-feature (orientations and color components) through a PCA on the corresponding set of scales. On the other hand, variability and richness of structural content have been proven as driving attention [17]. Therefore, we have chosen a measure of distance between local and global structure to represent distinctiveness. Once scales are decorrelated, we extract the statistical distance at each point as the Hotelling's T^2 statistic:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1N} \\ \vdots & \vdots & \vdots \\ x_{S1} & \dots & x_{SN} \end{pmatrix} \rightarrow (PCA) \rightarrow \mathbf{T}^2 = \left(T_1^2 \, , \, \cdots \, , \, T_N^2\right). \tag{4}$$

That is, being S the number of scales (original coordinates) and N the number of pixels (samples), we compute the statistical distance of each pixel (sample) in the decorrelated coordinates. Given the covariance matrix ($\mathbf{W}$), T^2 is defined as:

$$T_i^2 = (\mathbf{x_i} - \bar{\mathbf{x}})' \mathbf{W^{-1}} (\mathbf{x_i} - \bar{\mathbf{x}}). \tag{5}$$

This is the key point of our approach to the integration process. This multivariate measure of the distance from a feature vector associated to a point in the image, to the average feature vector of the global scene, is in fact, a measure of the local feature contrast [18].

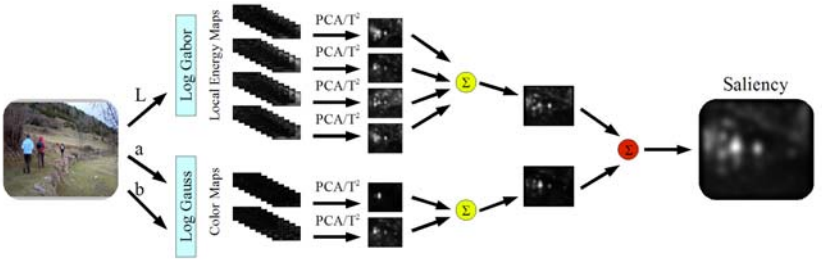

Fig. 1. Bottom-up Saliency Model

Final Map. The final saliency map is obtained normalizing to [0,1], smoothing, and summing the extracted maps, first within each feature dimension and next with the resulting local energy conspicuity and color conspicuity maps. In this way we obtain a unique measure of saliency for each point of the image.

Computational Complexity. The whole process involves two kinds of operations. Firstly, filtering for decomposition and smoothing, has been realized in the frequency domain, as the product of the transfer functions of the input and the filters, using the Fast Fourier Transform (FFT) and its inverse (IFFT). This implies a computational complexity of O(N log(N) + N), being N the number of pixels of the image. The other operation is PCA with a complexity of $O(S^3 + S^2 N)$, being S the number of scales (dimensionality) and N the number of pixels (samples) [19]. We are interested in the dependency on the number of pixels, being O(N), since the number of scales remains constant. Therefore, the overall complexity is given by O(N log(N)).

3 Experimental Results and Discussion

In this work we demonstrate the efficiency of the model predicting eye fixations in natural images. In Section 3.1 we show the performance of the model in terms of AUC values from ROC analysis, on a public image dataset. We compare these results with those obtained by other models representative of the state of the art. Moreover, we discuss the details of this procedure, as well as the difficulties and limitations that it poses. This last issue motivates Section 3.2, where we use a metric based on the K-L divergence.

3.1 ROC Values

In this experiment we employ an image dataset published by Bruce & Tsotsos. It is made up of 120 images, and the corresponding fixation data for 20 different subjects. A detailed description of the eye-tracking experiment can be found in [4].

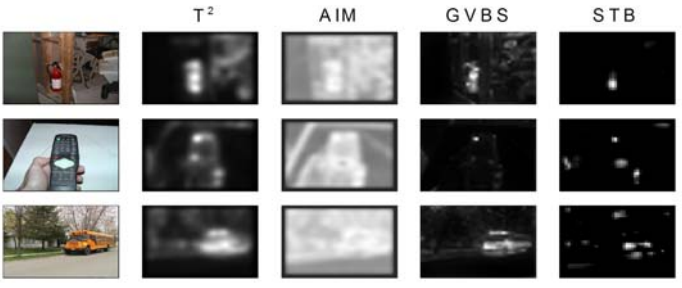

Fig. 2. Results of the model (T^2) with three of the images used. Results for other models, have been obtained with the corresponding open access version.

Table 1. AUC average values from computed like in [6]. (*published by the authors)

Model	AUC	std
T^2-Based	0.791	0.080
Gao et al. [6]*	0.769	---
GBVS [8]	0.688	0.119

The details of AUC computation from ROC analysis face a difficulty from the beginning. Bruce & Tsotsos construct one unique ROC curve for the complete set of images, with the corresponding AUC. The uncertainty is provided, based on the proposal of Cortes & Mohri [20]. They give a value for their model and the model of Itti et al. [12]. On the other hand Gao et al. compare these results with those obtained by their model, but with another procedure. They construct one ROC curve and extract the corresponding AUC for each image. They take the average of the AUC as the overall value, but they don't provide any estimation of uncertainty. The same averaging procedure is employed by Harel et al. but on a different set of images [8].

When computing AUC with the average procedure we find a problem: standard deviation is larger than the differences between models, although these differences can be also large. This is reflected in table 1. The model of Gao et al. should have a similar value of standard deviation, since partial graphical values are similar to the models of Bruce & Tsotsos and Itti et al. [6].

Instead, we can proceed like Bruce & Tsotsos, who obtain a much tighter value for a 95% uncertainty, while the AUC is very similar (slightly lower). Thus, this would make it possible to rank the models by their overall behavior on the whole dataset. Results are shown in table 2. Our model has equivalent results to Bruce & Tsotsos, improving the performance of all of the other models.

However, this approach hides a problem. We are analyzing all the scenes as a unique sample, instead of considering each scene separately. Hence, the approach of Bruce & Tsotsos means to loose the inter-scene variance, performing a global assessment. Then, a question arises: does the kind of scene affect the ranking of models?. That is, could there be scene-biased models? If the kind of scene is

Table 2. AUC computed from a unique ROC curve ([*]values provided by [5])

Model	AUC	std
T^2-Based	0.776	0.008
AIM [9][*]	0.781	0.008
Itti et al. [12][*]	0.728	0.008
GBVS [8]	0.675	0.008
STB [21]	0.569	0.011

important, probably this dataset is not representative enough of natural images. In fact, urban and man-made scenes are clearly predominant. For instance, there is no landscape, and there is only one image with an animal (but in an indoor environment).

This fact could help to explain the results reported by Harel et al. [8], that show a higher (and excellent) performance compared to models of Bruce & Tsotsos and Itti et al., using a different image dataset. We must notice here that these other images were gray-level (without color information), and were mainly images of plants.

3.2 K-L Divergence Values

In this Section we employ the K-L divergence to compare priority maps from fixations with saliency maps, similarly to [10]. As priority maps we use the density maps computed by Bruce & Tsotsos to reflect the foveated region with each fixation.

The priority map can be interpreted as a measure of the probability of each point to attract gaze, and the saliency map can be viewed, in turn, as a prediction of that probability. Hence, it makes sense to compare both distributions through the K-L divergence. It is worth noting that, instead of gray levels probabilities [5], we compare distributions of probabilities in the space.

Table 3. K-L comparison. Other models have been executed using their default values.

Model	K-L	std
T^2-Based	1.3	0.3
AIM [9]	1.7	0.3
GBVS [8]	2.1	0.6
STB [21]	13.0	1.6

With this aim, we obtain probability maps simply dividing a given map by the sum of its gray-level values. We denote by $h_i = h(x, y)$ and $m_i = m(x, y)$ the priority map from fixations and the saliency map from each model (taken as probability distributions) respectively. Then, being N the number of pixels, we compute the K-L divergences:

$$D_{KL}(h, m) = \sum_{i=1}^{N} h_i \cdot \log \frac{h_i}{m_i} \tag{6}$$

It is possible to use other procedures to construct the priority maps, to take into account other parameters like the duration of fixations, and not merely their positions [22]. Therefore, this metric should be viewed as an interesting complement to ROC analysis.

Nevertheless, the results shown in table 2, lead to a similar interpretation to that derived from analysis of ROC curves. Standard deviations are similar in relative order of magnitude. Our model exhibits an slightly better performance than the model of Bruce & Tsotsos [5], and again clearly better than the model proposed by Harel et al [8]. The implementation proposed by Walther [21] of the model of Itti & Koch [23] exhibits again the worst result.

4 Conclusions

In this work we have shown a simple and light model of saliency, that resorts to the decorrelation of the responses to a Gabor-like bank of filters. This mechanism is biologically plausible and could have an important role in the influence of NCRF when V1 cells are subjected to natural stimuli [9][16].

We have validated the model, comparing its performance with others, in the prediction of human fixations. Using ROC analysis we have obtained a result equivalent to that achieved by Bruce & Tsotsos [5], after the optimization of their decomposition process. With the same model in a previous work they obtained an AUC value of 0.7288 [4], clearly lower. On the other hand, using a different metric based on K-L divergence, that takes into account the area of foveation of each fixation, the model performs slightly better than the approach of Bruce & Tsotsos. Other models [6][8][21] deliver worse results in both comparisons.

Similarly to Bruce & Tsotsos [5], we avoid any parameterization of the process, beyond the initial decomposition of the image. However, this decomposition remains ordered and suitable for the incorporation, from the beginning, of top-down influences. Finally, our model of saliency presents a lower computational complexity than models that are benchmarks for psychophysical plausibility.

Acknowledgments. This work has been granted by the Ministry of Education and Science of Spain (project AVISTA TIN2006-08447), and the Government of Galicia (project PGIDIT07PXIB206028PR).

References

1. Tsotsos, J.K.: Computational foundations for attentive Processes. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) Neurobiology of Attention, pp. 3–7. Elsevier, Amsterdam (2005)
2. Frintrop, S., Jensfelt, P.: Attentional Landmarks and Active Gaze Control for Visual SLAM. IEEE Transactions on Robotics, Special Issue on Visual SLAM 24(5) (2008)
3. Harel, J., Koch, C.: On the Optimality of Spatial Attention for Object Detection, Attention in Cognitive Systems. In: WAPCV (2008)

4. Bruce, N., Tsotsos, J.K.: Saliency Based on Information Maximization. In: NIPS, vol. 18, pp. 155–162 (2006)
5. Bruce, N., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9(3), 1–24 (2009)
6. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. Journal of Vision 8(7), 13, 1–18 (2008)
7. Gao, D., Mahadevan, V., Vasconcelos, N.: The discriminant center-surround hypothesis for bottom-up saliency. In: NIPS (2007)
8. Harel, J., Koch, C., Perona, P.: Graph-Based Visual Saliency. In: NIPS, vol. 19, pp. 545–552 (2007)
9. Olshausen, B.A., Field, D.J.: How Close Are We to Understanding V1? Neural Computation 17, 1665–1699 (2005)
10. Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 802–817 (2006)
11. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Local energy variability as a generic measure of bottom-up salience. In: Yin, P.-Y. (ed.) Pattern Recognition Techniques, Technology and Applications, In-Teh, Vienna, pp. 1–24 (2008)
12. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
13. Field, D.J.: Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells. Journal of the Optical Society of America A 4(12), 2379–2394 (1987)
14. Kovesi, P.: Invariant Measures of Image Features from Phase Information. Ph.D. Thesis, The University or Western Australia (1996)
15. Morrone, M.C., Burr, D.C.: Feature Detection in Human Vision: A Phase-Dependent Energy Model. Proceedings of the Royal Society of London B 235, 221–245 (1988)
16. Vinje, W.E., Gallant, J.L.: Sparse coding and decorrelation in primary visual cortex during natural vision. Science 287, 1273–1276 (2000)
17. Zetzsche, C.: Natural Scene Statistics and Salient Visual Features. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) Neurobiology of Attention, pp. 226–232. Elsevier, Amsterdam (2005)
18. Nothdurft, H.C.: Salience of Feature Contrast. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) Neurobiology of Attention, pp. 233–239. Elsevier, Amsterdam (2005)
19. Sharma, A., Paliwal, K.K.: Fast principal component analysis using fixed-point algorithm. Pattern Recognition Letters 28, 1151–1155 (2007)
20. Cortes, C., Mohri, M.: Confidence intervals for the area under the ROC curve. In: NIPS, vol. 17, p. 305 (2005)
21. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407 (2006)
22. Ouerhani, N.: Visual Attention: Form Bio-Inspired Modelling to Real-Time Implementation, PhD thesis, University of Neuchatel (2004)
23. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40, 1489–1506 (2000)