A MODEL OF BOTTOM-UP VISUAL ATTENTION USING CORTICAL MAGNIFICATION

Ala Aboudib, Vincent Gripon and Gilles Coppin

Télécom Bretagne
Technopôle Brest-Iroise, CS 83818 – 29238 Brest cedex 3, France
Lab-STICC UMR CNRS 6285

ABSTRACT

The focus of visual attention has been argued to play a key role in object recognition. Many computational models of visual attention were proposed to estimate locations of eye fixations driven by bottom-up stimuli. Most of these models rely on pyramids consisting of multiple scaled versions of the visual scene. This design aims at capturing the fact that neural cells in higher visual areas tend to have larger receptive fields (RFs). On the other hand, very few models represent multi-scaling resulting from the eccentricity-dependent RF sizes within each visual layer, also known as the cortical magnification effect. In this paper, we demonstrate that using a cortical-magnification-like mechanism can lead to performant alternatives to pyramidal approaches in the context of attentional modeling. Moreover, we argue that introducing such a mechanism equips the proposed model with additional properties related to overt attention and distance-dependent saliency that are worth exploring.

Index Terms— cortical magnification, bottom-up attention, saliency, multi-scale

1. INTRODUCTION

The study of visual attention and its computational modeling is an emerging field. Object recognition is one of its numerous potential domains of applications [1]. The seminal work of Treisman and Gelad in the Feature Integration Theory (FIT) [2] investigated the key role of focusing attention in correctly associating visual properties related to the same object. Later on, Koch and Ullman [3] introduced the first theoretical model suggesting a possible neuro-inspired bottom-up mechanism for guiding attention. The first working implementation of Koch and Ullman's model was later proposed by Itti and Koch in [4] and became a landmark model for bottom-up attentional modelling.

Since then, many works aimed at dealing with the object recognition problem from an attentional perspective. Rybak [5] proposed an algorithm that associates a scanpath drawn from several fixations with the contents of fixated regions to perform learning and recognition. The role of gist information and context in guiding visual attention was elaborately studied by Torralba [6], leading to a better understanding of the different kinds of top-down influences on attention. Later on, Walther and Koch proposed a unified framework for combining attention and object recognition [7].

The main contribution of this paper is to demonstrate that cortical-magnification-like mechanisms can be applied as an alternative to pyramidal multi-scale processing traditionally used in computational models of visual attention. We introduce a new model and we compare it to the renowned Itti and Koch model originally proposed in [4]. We obtain similar or even better performance while using a much smaller number of feature maps. After that, we explore some interesting properties the magnification mechanism adds to the proposed model.

The rest of this paper is organized as follows: Section 2 introduces some related work. In section 3 we explain the technical details of the proposed model. We introduce and discuss simulation results in section 4. Section 5 is a conclusion.

2. RELATED WORK

Image pyramids are widely used tools in models of bottom-up visual attention. For instance, Itti and Koch in their famous work [4] used dyadic Gaussian pyramids to generate several spatial scales of each input image. Feature maps are then created by applying a cross-scale subtraction operator on these images.

A different method was proposed by Rybak in [5]. He modelled multiple spatial scales by creating a single retinal image centered at each fixation point. Each such image is composed of three concentric regions with increased blurring as they go far from the fixation center. Advani in his his MR-AIM model [8] creates multi-resolution images by concatenating patches extracted from a Gaussian pyramid. Multi-resolution processing was also used in many other models such as [9] and by using wavelet transforms in [10].

More recently, the computational modeling of cortical magnification began to emerge in a more formal way. One

This work was supported by the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n[Pleaseinsert"PrerenderUnicode–˝intopreamble] 290901.

example is the work of Freeman and Simoncelli on Metamers in the ventral stream [11]. Another example is the work of Poggio [12] on the computational role of cortical magnification as a sampling extension of the "Magic theory" [13] and its role in scale and shift invariance for pattern recognition [14].

3. METHODOLOGY

The proposed model is based on the original work of Itti and Koch in [4]. Figure 1 shows the basic architecture of the proposed model. Input is provided to the model as an RGB image M. It is then embedded into a square black background to form the RGB scene E. We do such embedding in order to force the input to be square-shaped for an easier manipulation. Then, a grey-scale version of the scene E_r is created as a simple average of the R,G and B channels of E.

A visual angle Θ_M is associated with the image M. This simulates the fact that viewing a given image from different distances changes the visual angles occupied by that image and thus might change the attentional behavior of the visual system. Another visual angle Θ_v is assigned to a central region of E that represents the fovea. Thus, the "further" the image is from the model's "eye", the smaller is the visual angle Θ_M occupied by the image M and the larger is the area that falls within the fovea. This is meant to simulate the case when an image is seen by a human subject from different distances.

We do not use Gaussian pyramids to generate multiple scales of the input scene as in [4]. We replace this step by generating feature maps using kernels with eccentricity-dependent sizes representing RFs. Eccentricity dependency is observed across several layers of the ventral stream including LGN, V1, V2, V4 and beyond [15][16]. This allows us to reduce the number of required feature maps from 42 in the model of Itti and Koch [4] to 9 maps in the proposed model while gaining in performance on predicting fixations (cf. Section 4).

3.1. Designing the kernel bank F

In order to create feature maps, a bank of spatial filters F is applied to the scene E and to its gray-scale version E_r. Filter kernels in F are designed to emulate the eccentricity-dependency of RFs in LGN and V1. This dependency is behind what is known as the cortical magnification effect. Therefore, kernels in the center of our model's "visual field" MVF should be more numerous and smaller in size and become sparser with a linear increase in their diameters as they approach the periphery [15] [17]. To approximate this effect, we determine the locations of kernel centers and their sizes using a geometry very similar to the foveal Cartesian geometry proposed by [18] as explained below.

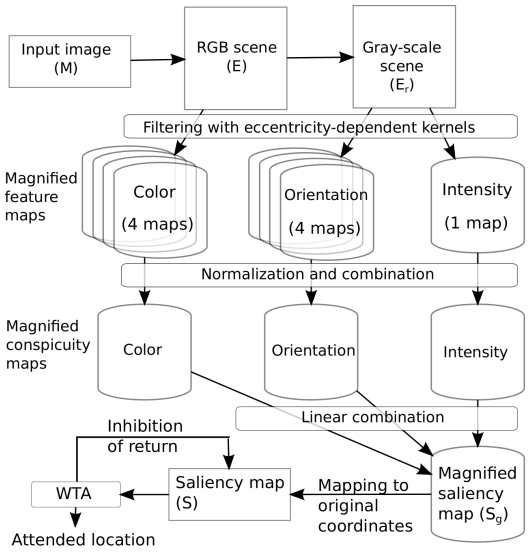


Fig. 1: Model architecture.

A square 2D map with the same height and width as the scene E is created as in the illustrative Figure 2(left). A central square region is chosen to represent the fovea containing a fixed number of kernel centers. Then, a number of concentric square rings each made of equally-spaced kernel centers on each side is created for the periphery. The side of each ring contains two more centers than its direct preceding inner ring. This allows these concentric rings to fit into a smaller square map which has as many pixels as there are kernels in F.(cf. Figure 2(right)). We refer to the latter as a "magnified map" and a subscript g is associated to its name. When a given kernel is applied to the scene by an inner product, the resulting value is assigned to the corresponding pixel in the magnified map.

All kernels whose centers are within the the model's fovea have the same radius. Similarly, all periphery kernels whose centers belong to the same square have the same radius. These radii are calculated from the linear equation $r = ax$ where x is the eccentricity (in visual angles) of the corner pixels of the square to which a kernel belongs and a is the inverse magnification factor M^{-1}.

The eccentricity of corner pixels of the fovea are determined from the visual angle Θ_v assigned to the fovea. Eccentricities of corner pixels of periphery rings are determined in such a way that an overlap p occurs between kernels of F along the diagonal.

3.2. Magnified feature maps

Nine feature maps are directly calculated; one map for the intensity modality, 4 maps for the color modality and 4 maps for the orientation modality.

Only one magnified intensity feature map I_g is created from the gray-scale scene E_r. A filter bank F_{DoG} is used

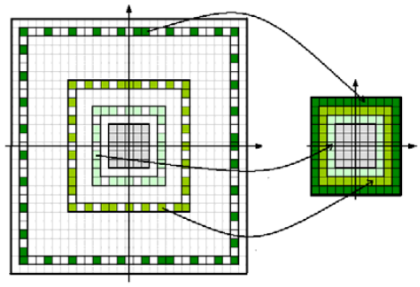


Fig. 2: Placement of kernel centers (left) and mapping to a magnified map (right). Figure adapted from [18].

with L2-normalized Difference of Gaussian (DoG) kernels with zero mean to approximate the center-surround functionality of RFs in LGN. Then we apply a special convolution operator $\circledast_r$ between the gray-scale scene E_r and the filter bank F_{DoG}:

$$I_g = E_r \circledast_r F_{DoG} \tag{1}$$

The operator $\circledast_r$ consists of a rectified normalized inner product between each kernel in F_{DoG} and the patch it spans in E_r. The scalar value resulting from the inner product then occupies one pixel in the magnified map I_g as explained in Section 3.1.

Four magnified orientation feature maps $O_g(\theta)$ with $\theta \in \{0°, 45°, 90°, 135°\}$ are created. Then the same convolution operator is applied:

$$O_g(\theta) = E_r \circledast_r F_{O(\theta)} \tag{2}$$

where $F_{O(\theta)}$ is a filter bank F with L2-normalized Gabor kernels [19] with orientation θ and zero mean. Each Gabor filter $f_{O(\theta)} \in F_{O(\theta)}$ is characterized by its orientation θ, effective width σ, wavelength λ and its aspect ratio γ.

Finally, we use F_{DoG} and the RGB scene E to create four magnified feature maps for the color-opponent channels $\mathcal{RG}_g$, $\mathcal{GR}_g$, $\mathcal{BY}_g$, $\mathcal{YB}_g$ as follows:

$$\mathcal{RG}_g = E \circledast_{rg} F_{DoG} \tag{3}$$
$$\mathcal{GR}_g = E \circledast_{gr} F_{DoG} \tag{4}$$
$$\mathcal{BY}_g = E \circledast_{by} F_{DoG} \tag{5}$$
$$\mathcal{YB}_g = E \circledast_{yb} F_{DoG} \tag{6}$$

Each of the convolution operators $\circledast_{rg}$, $\circledast_{gr}$, $\circledast_{by}$ and $\circledast_{yb}$ consists of a half-wave rectified (negative values are set to zero) normalized inner product between each kernel in F_{DoG} and a corresponding 2D patch t. In the case of $\circledast_{rg}$, the patch t is a concatenation between the area of the R component of E spanned by the center of the kernel and the area of the G component corresponding to the surround. The inverse holds for $\circledast_{gr}$. Similarly, when applying the operator $\circledast_{by}$, the patch t becomes a concatenation between the center area of the B component with the surround area of the yellow component Y (the average of the red and green components). The inverse hold for $\circledast_{yb}$. This transformation is inspired by the DKL color space.

It should be noted here that kernel filters that overlap with the image M borders embedded in the scene E or E_r could sometimes give high saliencies. This is most likely a result of the saliency of the image relative to the black background in which it is embedded. So all such values are set to zeros in all feature maps.

3.3. Magnified conspicuity maps and the saliency map

Three magnified conspicuity maps $\bar{I}_g$, $\bar{O}_g$ and $\bar{C}_g$ are created for intensity, orientation and color, respectively. The process is as follows:

$$\bar{I}_g = \mathcal{N}(I_g) \tag{7}$$

$$\bar{O}_g = \mathcal{N}\left(\sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}\left(O_g(\theta)\right)\right) \tag{8}$$

$$\bar{C}_g = \mathcal{N}\left[\mathcal{N}(|\mathcal{RG}_g - \mathcal{GR}_g|) + \mathcal{N}(|\mathcal{BY}_g - \mathcal{YB}_g|)\right] \tag{9}$$

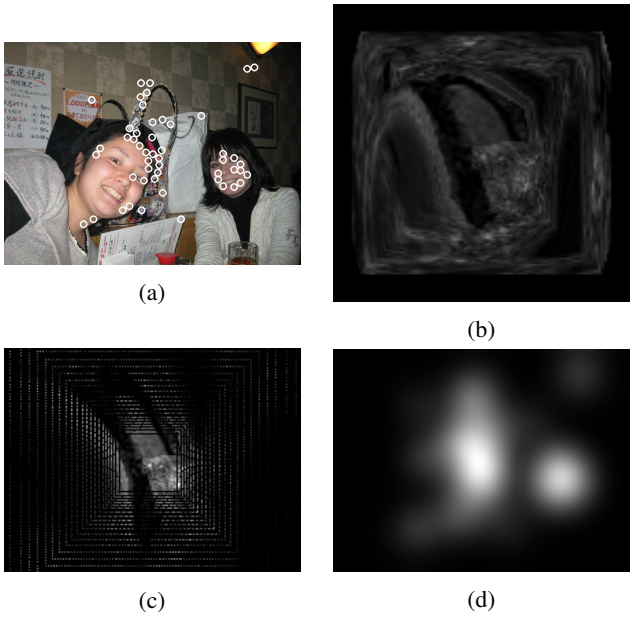
where $\mathcal{N}(.)$ is the normalization operator designed by Itti and Koch in [4]. The global magnified saliency map is then calculated as a linear combination of conspicuity maps:

$$S_g = \frac{1}{3}(\bar{I}_g + \bar{O}_g + \bar{C}_g) \tag{10}$$

Then a simple inverse mapping of pixels in S_g into a map with the same size as the scene E is done. After this, the black background is removed to get a standard saliency map S that has the same size as the input image M. Notice that S is a very sparse map composed of a central fovea surrounded by concentric square rings. In order to be able to compare to other models, a continuous saliency map is created by convolving a Gaussian kernel on the most salient locations extracted from S using an alternating Winner-Take-All (WTA) and Inhibition of Return mechanism. See Figure 3.

4. RESULTS AND DISCUSSION

We used the Judd Benchmark [20] for partially optimizing some parameters such as the number of fixations extracted, the width of the Gaussian kernel convolved on fixation points to create the final saliency map and the value of Θ_M. Then, we ran the model on the MIT Saliency Benchmark [21] which contains 300 natural color images. We set the visual angle $\Theta_M = 50°$ for each input image M and $\Theta_v = 1°$ for the fovea which is close to its actual size in the human eye. The number of kernels in a bank F corresponding to the fovea is set to 41x41 units which is close to the number of units in the foveola of V1 estimated in [12]. The overlap parameter p is



(a) (b) (c) (d)

Fig. 3: Fixations generated by the proposed model and the corresponding saliency maps: (a) Image M (from the MIT Saliency Benchmark) with the first 50 fixations (b) The magnified saliency map S_g (c) A cropped region of the saliency map S after inverse mapping from S_g showing the fovea and some periphery square rings (d) The continuous saliency map made with a Gaussian convolved on fixation locations.

set to 0.8. The slope of the inverse magnification curve presented in Section 3.1 is set to $a = 0.16$ as suggested by Gattass in [17] for V1. Parameters of Gabor kernels are inspired by [22]; we set $\gamma = 0.3$, $\sigma/\lambda = 0.8$ and $d/\sigma = 2.5$ where d is the diameter of a given kernel. Parameters of center-surround DoG kernels were adapted from [23]; we set $d/\sigma_c = 11.8$, $\sigma_s/\sigma_c = 3$ and $g_s/g_c = 0.8$ where σ_c and σ_s are the standard deviations of the center and surround Gaussians of a given DoG kernel, respectively. While g_c and g_s are the strengths of the center and surround Gaussians, respectively.

Table 1 shows the scores of our models according to several metrics used by the benchmark and how they are compared to the Itti and Koch model [4]. Our results and comparisons with many more models are also available on the MIT Saliency Benchmark website http://saliency.mit.edu.

For all metrics, the proposed model performes much better than IttiKoch1 model which is the original implementation of [4] by its authors. On the other hand, the IttiKoch2 implementation of the same model by Jonathan Harel as a part of his GBVS toolbox gives a roughly similar performance; It has a similar performance for the similarity metric and higher AUC (Area Under the ROC curve) metrics. However, the proposed model has a better Correlation-Coefficient (CC), Normalized Scanpath Saliency (NSS) and Earth Mover's Distance (EMD).

Metric	**Our model**	**IttiKoch1**	**IttiKoch2**
Similarity	0.44	0.20	0.44
AUC (Judd)	0.74	0.60	0.75
AUC (Borji)	0.72	0.54	0.74
Shuffled AUC	0.58	0.53	0.63
CC	0.39	0.14	0.37
NSS	0.99	0.43	0.97
EMD	4.24	5.17	4.26

Table 1: Comparison between the performance of the proposed model and Itti and Kock model on the MIT Saliency Benchmark.

It is worth noting that the IttiKoch2 implementation and many other models are optimized for blur and center-bias. The proposed model, however, has no explicit center-bias applied and only a minor blur optimization. We hypothesize that an effect similar to center-bias arise naturally in our model due the magnification factor. This is more biologically plausible than explicitly applying a Gaussian mask to account for center-bias.

5. CONCLUSION

In this paper, we introduced a new model of bottom-up visual attention based on Itti and Koch work in [4]. We demonstrated that using cortical magnification could be an alternative to pyramidal approaches to multi-scale image processing for modeling attention. We also showed that this allows us to boost performance while using a much smaller number of feature maps which can automatically account for center-bias phenomena.

The proposed model has also several other properties discussed below. They are worth investigating in future works.

Associating a visual angle Θ_M to input images allows us to study the effect of the distance from which an image is displayed on saliency. This is an interesting effect that is not often considered by saliency models or benchmarks. We suggest that distance-dependent saliency should be evaluated by future benchmarks.

The proposed model is inherently adapted to performing overt attention where the position of the fovea moves across the scene. This is not directly the case for most attentional models. Given the magnification property, different saliency maps of the same scene are generated each time the fovea moves. This could have important implications on rethinking the role of the inhibition of return.

Finally, kernels with different sizes provide different information. Larger kernels provide more global information about the scene. By exploiting this property and using kernels as inputs to associative memories, one can imagine a unified framework where attention, gist information and object recognition techniques can all interact together.

6. REFERENCES

[1] Ali Borji and Laurent Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.

[2] Anne M Treisman and Garry Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[3] Christof Koch and Shimon Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*, pp. 115–141. Springer, 1987.

[4] Laurent Itti, Christof Koch, and Ernst 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.

[5] IA Rybak, VI Gusakova, AV Golovan, LN Podladchikova, and NA Shevtsova, "A model of attention-guided visual perception and recognition," *Vision research*, vol. 38, no. 15, pp. 2387–2400, 1998.

[6] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.," *Psychological review*, vol. 113, no. 4, pp. 766, 2006.

[7] Dirk B Walther and Christof Koch, "Attention in hierarchical models of object recognition," *Progress in brain research*, vol. 165, pp. 57–78, 2007.

[8] Siddharth Advani, John Sustersic, Kevin Irick, and Vijaykrishnan Narayanan, "A multi-resolution saliency framework to drive foveation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2596–2600.

[9] Rebeca Marfil, Antonio J Palomino, and Antonio Bandera, "Combining segmentation and attention: a new foveal attention model," *Frontiers in computational neuroscience*, vol. 8, 2014.

[10] ZhiQiang Li, Tao Fang, and Hong Huo, "A saliency model based on wavelet transform and visual attention," *Science China Information Sciences*, vol. 53, no. 4, pp. 738–751, 2010.

[11] Jeremy Freeman and Eero P Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[12] Tomaso Poggio, Jim Mutch, and Leyla Isik, "Computational role of eccentricity dependent cortical magnification," *arXiv preprint arXiv:1406.1770*, 2014.

[13] F Anselmi, JZ Leibo, L Rosasco, J Mutch, A Tacchetti, and T Poggio, "Magic materials: a theory of deep hierarchical architectures for learning sensory representations," *CBCL paper*, 2013.

[14] Leyla Isik, Joel Z Leibo, Jim Mutch, Sang Wan Lee, and Tomaso Poggio, "A hierarchical model of peripheral vision," 2011.

[15] R Gattass, CG Gross, and JH Sandell, "Visual topography of v2 in the macaque," *Journal of Comparative Neurology*, vol. 201, no. 4, pp. 519–539, 1981.

[16] David H Hubel and Torsten N Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[17] R Gattass, AP Sousa, and CG Gross, "Visuotopic organization and extent of v3 and v4 of the macaque," *The Journal of neuroscience*, vol. 8, no. 6, pp. 1831–1845, 1988.

[18] José Martínez and Leopoldo Altamirano Robles, "A new foveal cartesian geometry approach used for object tracking.," *SPPRA*, vol. 6, pp. 133–139, 2006.

[19] Dennis Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[20] Tilke Judd, Frédo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[21] Zoya Bylinskii, Tilke Judd, Frédo Durand, Aude Oliva, and Antonio Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.

[22] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, "Robust object recognition with cortex-like mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 411–426, 2007.

[23] Robert W Rodieck, "Quantitative analysis of cat retinal ganglion cell response to visual stimuli," *Vision research*, vol. 5, no. 12, pp. 583–601, 1965.