A Biologically Inspired Framework for Visual Information Processing and an Application on Modeling Bottom-Up Visual...

Article *in* Cognitive Computation · September 2016

DOI: 10.1007/s12559-016-9430-8

CITATIONS
0

READS
99

3 authors:

Ala Aboudib
Institut Mines-Télécom
7 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Vincent Gripon
Institut Mines-Télécom
74 PUBLICATIONS **398** CITATIONS
SEE PROFILE

Gilles Coppin
Telecom Bretagne / Lab-STICC UMR 6285
82 PUBLICATIONS **167** CITATIONS
SEE PROFILE

A biologically-inspired framework for visual information processing and an application on modeling bottom-up visual attention

Ala Aboudib · Vincent Gripon · Gilles Coppin

Abstract An emerging trend in visual information processing is toward incorporating some interesting properties of the ventral stream in order to account for some limitations of machine learning algorithms. Selective attention and cortical magnification are two such important phenomena that have been the subject of a large body of research in recent years. In this paper, we propose a new framework for visual information acquisition and representation that emulates the architecture of the primate visual system by integrating features such as retinal sampling and cortical magnification while avoiding spatial deformations and other side effects produced by current models that tried to implement these two features. It also explicitly integrates the notion of visual angle, which is rarely taken into account by vision models. We argue that this framework can provide the infrastructure for implementing vision tasks such as object recognition and computational visual attention algorithms. It also raises important questions about the role of the newly integrated features on vision behavior. Moreover, we propose an algorithm for bottom-up visual attention implemented using the proposed framework, and show that it can attain state-of-the-art performance, and provide a better insight on the significance of studying the role of the visual angle more closely.

Keywords cortical magnification · bottom-up attention · saliency · retinal sampling · foveation

This work was supported by the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n° 290901.

Télécom Bretagne
Technopôle Brest-Iroise, CS 83818 – 29238 Brest cedex 3, France
Lab-STICC UMR CNRS 6285
Email addresses: {name}.{surname}@telecom-bretagne.eu

1 Introduction

Vision and the visual system have been an active area of research for many centuries. Interest in exploring this territory has been motivated by a wide variety of applications. Ophthalmology was one of the first domains to benefit from such discoveries. More recently, that interest has been widely driven by the desire to learn more about the brain and decipher its neural code. A better understanding of the neural code has enabled to design better machine learning algorithms for computer vision, and for artificial intelligence in a more general sense.

The discovery of simple and complex cells in the famous work by Hubel and Wiesel on receptive fields in the cat's visual cortex (Hubel and Wiesel, 1959) marked a new era in vision research. It revolutionized the way the visual system is studied and understood, and allowed for the emergence of 'computational neuroscience', a new field founded by David Marr whose theory on vision is still very influential (Marr, 1982).

More recently, deep learning networks have achieved an unprecedented performance on many visual tasks such as image categorization (Krizhevsky et al, 2012). The architecture of these networks has been inspired by the multi-layered structure of the visual system and the hierarchical organization of simple and complex cells.

Some criticism of deep learning includes its limited performance on tasks such as unsupervised object discovery and localization, multiple instance recognition (MIL) (Zhu et al, 2015; Ray et al, 2010), recognizing spatial relationships between objects and its limited ability to generalizing to variable-scale representations of the learned classes without increasing the size of the training set (Lake et al, 2015). Another important problem of deep learning, according to (Ranzato et al, 2015), is its computational cost, which renders it impractical for very high resolution images. This called some researchers to get a closer look at the visual system and some of its overlooked properties to address these limitations.

One such property is selective visual attention that guides covert processing biases and saccadic eyes movements. The study of this property is an emerging trend in visual information processing. It finds its root in Treisman's Feature Integration Theory (FIT) (Treisman and Gelade, 1980). This theory provided a strong evidence of the fundamental role of attention for object recognition. This role was later explored by many researchers including (Koch and Ullman, 1987; Itti et al, 1998; Walther et al, 2004; Bonaiuto and Itti, 2005; Borji et al, 2014). It also motivated the recent emergence of attention-based recognition as in (Larochelle and Hinton, 2010; Zheng et al, 2015).

Cortical magnification is another ubiquitous feature of the visual system (Gattass et al, 1981, 1988). In addition to its role in reducing the amount of visual information entering the brain, Poggio has proposed that it might be a key property for enabling scale-invariant learning of objects (Isik et al, 2011; Anselmi et al, 2015).

In this paper, we propose a new framework for visual information acquisition that integrates these important features of the ventral stream. Our contributions are the following:

1. Introducing a new bio-inspired framework for visual information acquisition and representation that offers the following properties:
   - Providing a method for taking the distance between an image and the viewer into account. This is done by incorporating a visual angle parameter which is ignored by most visual acquisition models.
   - Reducing the amount of visual information acquired by introducing a new scheme for emulating retinal sampling and the cortical magnification effects observed in the ventral stream.
2. Providing a concrete application of the proposed framework by using it as a substrate for building a new saliency-based visual attention model, which is shown to attain state-of-the-art performance on the MIT saliency benchmark (Borji et al, 2013a).
3. Providing an online Git repository that implements the introduced framework that is meant to be developed as a scalable, collaborative project.

The rest of this paper is organized as follows: In Section 2, related work is reviewed. Section 3 introduces a brief anatomy of the visual system and its function, paving the way to Section 4 where a new vision framework is proposed that captures some interesting properties of the visual system. In Section 5, a new model of visual attention is proposed using the proposed framework. We show in Section 6 that coupling the proposed vision framework with the proposed attention model gives interesting results that motivates the utility of the framework. A discussion of some of the model's properties is also discussed in Section 6. Section 7 is a conclusion.

2 Related Work

Computational and mathematical modeling of the visual system have been the focus of many works in literature in recent years. The scope of such models includes mathematical models of single neurons (McCulloch and Pitts, 1943), neural assemblies through sparse coding techniques (Lee et al, 2006; Liu et al, 2015) and receptive field models (Rodieck, 1965; Marčelja, 1980) or even modeling complete visual layers, especially the retina (Wohrer and Kornprobst, 2009) or successions of layers representing early areas of the visual cortex such as in the Hmax model (Serre et al, 2007) or in the one proposed by David Marr in his famous work on vision (Marr, 1982).

Most vision models are designed to accomplish a specific task. For instance, the Hmax model is a view-based object recognition processor. It is inspired by the description of simple and complex cells in the primary visual cortex by Hubel and Wiesel (Hubel and Wiesel, 1962). A similar model was proposed in (LeCun et al, 1998), which also provides an implementation of simple and complex cells. However, it uses supervised learning coupled with

back-propagation to learn mathematical models of simple cells instead of fixing them beforehand. This allowed for unprecedented performance on many image classification tasks (Krizhevsky et al, 2012).

Some models have more general objectives. The virtual retina model in (Wohrer and Kornprobst, 2009) was proposed as a tool for researchers in neuroscience and neurophysiology to test their ideas and theories about visual function. Similarly, Walther and Koch proposed their model in (Walther and Koch, 2007) as a unified framework for implementing saliency-based visual attention and object recognition algorithms.

Although vision models are very numerous in literature, some important and even ubiquitous properties of the visual system are still absent in most of them. Retinal sampling and cortical magnification are examples of such properties. Poggio has argued that cortical magnification might play a fundamental role in introducing scale invariance in recognition (Poggio et al, 2014). However, a few models have used foveal-like transformations as an approximation to the cortical magnification effect (Rybak et al, 1998; Isik et al, 2011). While this imitates magnification in the sense that foveal and parafoveal regions are modeled at a higher resolution than the periphery, they differ in that the number of pixels representing the periphery is the same as in the original image, so the number of input pixels is not reduced (see Figure 1(b)).

At every layer of the visual system, an image zone that falls within the fovea is represented by more neurons than a zone with the same size falling within the periphery. One known method for emulating this is the log-polar representation (Schwartz, 1984). This method emulates retinal sampling very well by using a log-polar grid for sampling pixels of a given image. It then maps sampled pixels onto a rectangular-shaped image that has the drawback of having severe spatial deformations as shown in Figure 1(c). While this deformed representation has the advantage of being invariant to certain rotation and scale transformations, it is difficult to use such images for subsequent spatial processing used in many models such as the Hmax.

A different retinal sampling method that attempted to avoid log-polar-style deformation was proposed by (Martínez and Robles, 2006). It used sampling points organized in concentric squares to sample an image. These points can then perfectly fit into a square-shaped 2D array like in Figure 2. While this representation causes less deformation than the log-polar method, it still contains geometrical deformations along its diagonals as shown in Figure 1(d).

Thus, most known methods for generating retinal images have one of two major drawbacks. The first drawback is constraining the size of the retinal image to be equal to that of the input image, such as the Gaussian blurring method in Figure 1(b). This dependency of the output image size on the input is not observed in the visual system where the number of photo-receptors does not depend on the number of image pixels. Moreover, one important property of retinal sampling that such methods do not exploit is the fact that having a constant number of photo-receptors fixes an upper bound on the amount of information allowed to enter the visual system. The second drawback is the

(a) Original image.

(b) Blurring.

(c) Log-polar.

(d) Square-sampling.

Fig. 1: Some of the classical methods traditionally used for emulating retinal sampling and cortical images. Notice that blurring in (b) keeps the same number of pixel as in the original image. The log-polar and the square-sampling methods in (c) and (d) introduce severe spatial deformations that make further spatial filtering more challenging.

deformation introduced by methods that try to avoid the first drawback as in Figures 1(d) and (c).

We think that the main reason why such methods always have one of the above drawbacks is that they are constrained to producing an output image with a 'regular' shape. The term 'regular' here means a circular or a rectangular shape. This constraint is set such that the output image is suitable for presentation to a human observer or to be compatible with available image processing tools.

In the vision framework we propose in Section 4, no such shape constraints are fixed. Hence, we introduce a simple method for applying cortical magnification and retinal sampling in which the output is completely independent from the size of the input image without producing any geometrical deformations.

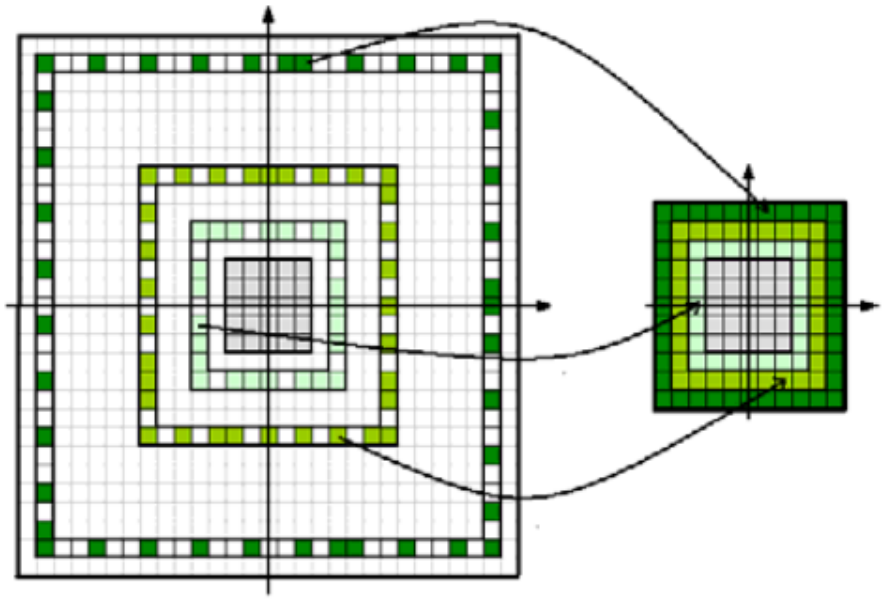


Fig. 2: Retinal sampling is emulated by using a set of points organized in concentric squares (left). Sampled points can then fit into a smaller square-shaped image (right) that represent the retinal image. (Martínez and Robles, 2006).

3 The visual system

3.1 Structure

The mammalian visual system is a highly intricate structure that exhibits a high level of organization. It starts at the eyes where light is trapped and transduced into neural signals. These signals are then conveyed through the optic nerve to the rest of the nervous system.

The first neural layer that processes visual signals is the retina. Although situated within the eyeball, the retina is considered as an integrate part of the brain. The retina is roughly composed of a three-layered feed-forward structure. The first layer is the Outer Nuclear Layer (ONL) containing photo-receptors. These photo-receptors synapse onto the bipolar cells that are found in the Inner Nuclear Layer (INL). Bipolar cells send their signals to ganglion cells in the Ganglion Cell Layer (GCL). An elaborate network of lateral connections are found between photo-receptors and are mediated by horizontal cells, while amacrine cells mediate lateral connections between bipolar cells (Dowling, 1987).

The optic nerve that is made of ganglion cell axons is the sole output of the retina. Most of the axons of the optic nerve project to the lateral geniculate

nucleus (LGN) in the thalamus. An important part of axons in the optic nerve also project to the superior colliculus. The optic nerve is the first stream where visual signals take the form of action potentials. Action potentials conveyed to the LGN by the optic nerve continue their way through the optic radiation which is another axonal structure. The optic radiation projects to the primary visual cortex V1 in the occipital lobe (Hubel and Wiesel, 1962).

The occipital lobe is divided into two distinct layers called V1 and V2. The optic radiation coming from LGN terminates in V1. At this point and starting from V2, the visual stream starts to diverge into to two distinct pathways: the ventral and the dorsal pathways. It has been argued that these pathways act as two independent visual systems with distinct functions (Goodale and Milner, 1992): the 'What' and the 'Where' systems. The 'What' system is situated in the temporal lobe. It is responsible for visual recognition tasks such as recognizing the identity of faces and other objects. In this system, higher visual areas such as V4, PIT, CIT and AIT are found. The 'Where' system, sometimes called the 'How' system, is found in the parietal lobe. It has been suggested that visually-guided behavior, such as reaching and grasping, is among the main functions of this system. Higher visual areas such as MT, LIP, MST and VIP are parts of this system. However, A later work by (Milner and Goodale, 2008) suggested the existence of a more complex interaction scheme than a simple separation into two independent systems.

It is worth pointing out that, in addition to the feed-forward pathway of axons, a rich feedback stream also go down throughout all the stages described so far.

3.2 Function

Two major families of photo-receptors are found in ONL: rods and cones. Rods are sensitive to low light conditions and are mainly responsible for night vision. On the other hand, cones are less sensitive to light, which makes them more adapted to day vision when light is abundant. Rods' and cones' main function is to transduce incoming photons into neural signals. These signals are further processed by the network of horizontal, bipolar and amacrine cells. They finally arrive at ganglion cells which translate them into action potentials and send them via the optic nerve to other areas.

There are two major types of ganglion cells with distinct functions, Parasol and midget cells. Parasol cells, also called M,Y or β cells, have wider receptive fields (RFs). They are characterized by a lower spatial resolution and a transient response to persistent stimuli. They are associated with achromatic vision. On the other hand, midget cells, which are sometimes called P, X or α cells, have smaller RFs. They have a higher spatial resolution and a lower temporal resolution than parasol cells, and they are associated with color vision. Each type of the above ganglion cells is divided into two sub-types called ON or OFF cells, which have complementary response levels. Hence, we find

parasol-ON, parasol-OFF, midget-On and midget-OFF cells (Salin and Bullier, 1995; Hubel and Wiesel, 1959).

Most ganglion cells are known for their center-surround configuration. ON ganglion cells are excited by the onset of light stimuli in their central region and inhibited by light in their surround region. The inverse holds for OFF ganglion cells. Rodieck was the first to propose an elegant mathematical model for spatial and temporal responses of ganglion cells in the form a difference of Gaussians (DoG) (Rodieck, 1965). This center-surround model is also involved in color vision. For example, some midget-ON cells encode the degree of red in their RFs; their center is excited by long wave (red) light and their surround is inhibited by medium wave (green) light. Another type is sensitive to blue, having a center excited by short wave (blue) light and a surround inhibited by medium and long wave light.

Neurons in higher visual areas respond to progressively more complex stimulus patterns. In the primary visual cortex V1, for example, neurons are tuned to simple oriented contours and spatial frequencies. The response of V1 cells, also called simple cells by (Hubel and Wiesel, 1959), are typically modeled mathematically by a Gabor filtering process, which consists in convoluting an image with a Gabor kernel made of the product of a 2D Gaussian kernel by a 2D cosine grating. Some neurons in V2 respond to stimuli as simple as oriented edges, but they are also tuned to illusory edges and a slightly more complex shapes. However, the complexity of spatial and temporal response patterns of neurons grows in complexity in higher visual areas.

3.3 Information reduction

The visual field (VF) of a single eye spans about 160° horizontally and 174° vertically. While a tremendous amount of visual information could be extracted from such a wide span, the visual system uses intelligent tricks to reduce the amount of acquired information. This reduction starts as early as the ONL layer in the retina; the visual field is sampled by the photo-receptors in a non-uniform fashion. The density of cones is very high in the central region of the retina called the fovea, spanning about 1°, and decreases logarithmically toward the periphery as shown in Figure 3. This distribution leads to what is called 'retinal sampling' (Salin and Bullier, 1995; Hubel and Wiesel, 1959).

The reduction of visual information by means of a 'privileged' fovea continues in subsequent areas of the visual system. For example, among ganglion cells of the same type, those which pool their inputs from photo-receptors near the fovea have smaller receptive fields than cells pooling their inputs from photo-receptors in the periphery. This phenomenon is known as the cortical magnification effect. This effect is also observed in LGN, V1, V2, V4 and even in higher visual areas. The ratio between the diameter of a given RF and its eccentricity stays relatively constant in a given visual layer and increases in higher areas as shown in Figure 4.

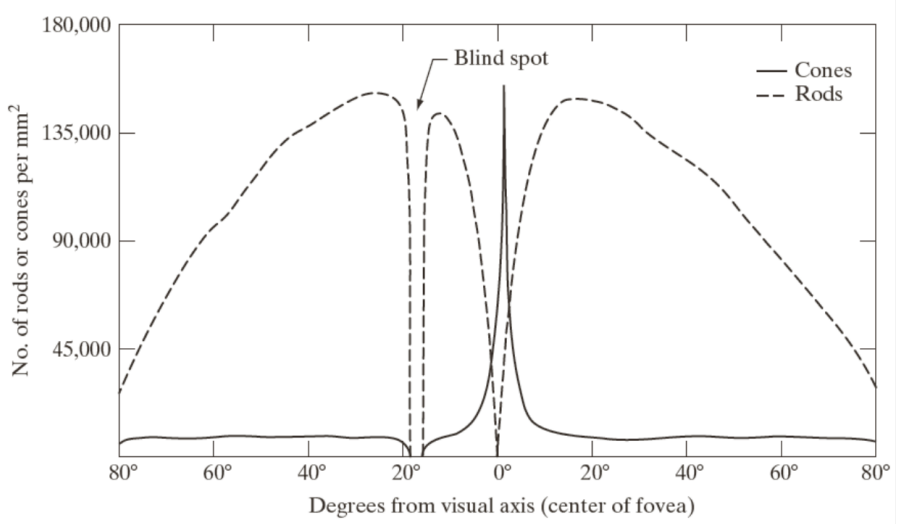


Fig. 3: Spatial density distribution of rods and cones in the retina (Gonzalez and Woods, 2002).

4 The proposed vision framework

In this section, we propose a model for visual information acquisition and representation in early layers of the visual system. This model is meant to be used as a framework for implementing visual processing tasks such as visual attention modeling presented in section 5, or object recognition algorithms that need hierarchical information processing. This model imitates the information reduction property of the visual system described in Section 3. It does this by emulating retinal sampling and the cortical magnification effect. This leads to some interesting properties discussed later in Section 6.

As we have seen in section 3, the early visual system can be functionally viewed as an arrangement of consequent layers. It starts at the photo-receptor layer (ONL) in the retina and continues through the GCL layer, the LGN, V1, V2 and so on. The transition from one layer into another can be viewed as a mapping mediated by synaptic connections constituting receptive fields.

The model we propose is made of two basic components: visual layers modeled as point clouds, and mapping functions between these layers in the feed-forward direction.

4.1 Notation

In this paper, polar coordinates and their corresponding Cartesian coordinates are sometimes used interchangeably depending on the context. The radial component of a polar coordinate is always denoted by the roman letter r, while

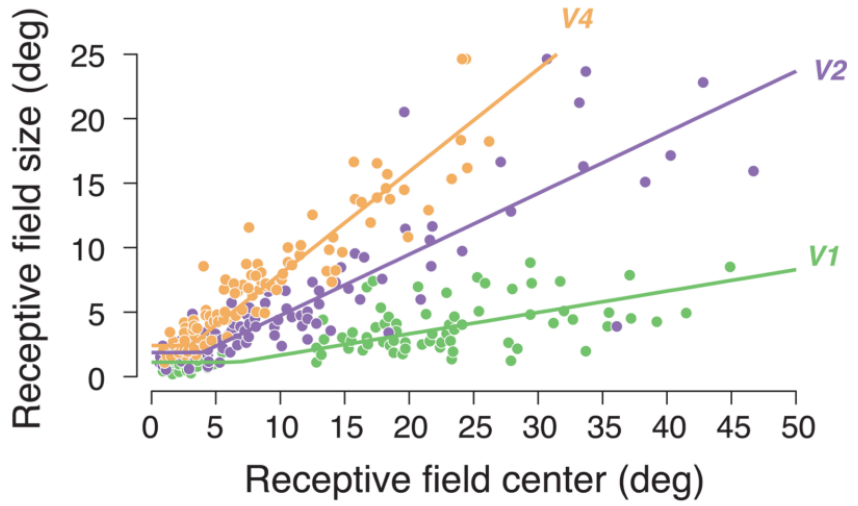


Fig. 4: Cortical magnification factors in V1, V2 and V4 adapted from (Gattass et al, 1981) and (Gattass et al, 1988) by (Freeman and Simoncelli, 2011).

the angle is denoted by the Greek letter ω. The Cartesian version of (r, ω) is always denoted by (x, y), where $x = r\cos\omega$ and $y = r\sin\omega$. If polar coordinates are written with super- and/or subscripts, these same super- and/or subscripts are attached to their Cartesian versions, and vice versa.

4.2 A generic model for visual layers

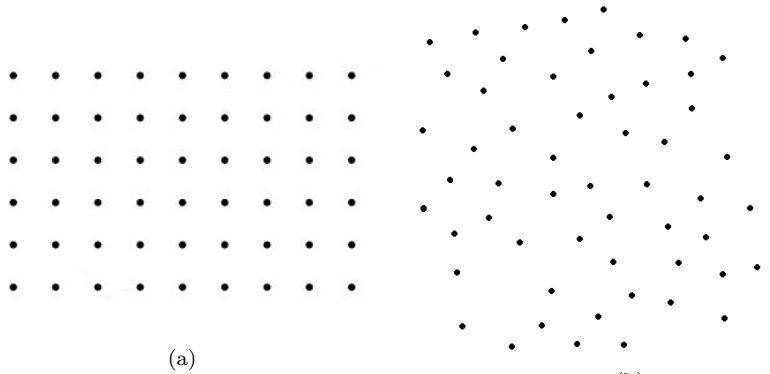
Visual layers as well as input stimuli are modeled as point clouds using a set representation. This representation can be used to instantiate any number of layers, which is a variable parameter between vision models, by providing a generic description that captures common properties of layers in the visual system as well as input stimuli, such as the visual field spanned by a layer, its fovea size, spatial distribution of cells and the distribution of associated receptive fields.

However, this representation focuses on two main properties of the visual system. First, it is adapted to implementing the information reduction properties in the form of retinal sampling and cortical magnification without introducing any deformations. Second, it implements the notion of visual angle which determines the visual field span associated with a given layer. The latter property is one main difference between the vision framework we propose and the one proposed in (Walther and Koch, 2007).

Hence, the structure of a given visual layer can be captured by our model using the following generic definition:

$$\mathcal{C}(\Theta^c, \psi^c, \mathcal{D}^c) = \{f_k^c | f_k^c : \mathbb{R}^2 \rightarrow \mathbb{R}, \\ \sigma(\text{diam}(\text{dom}(f_k^c))) = \Theta^c, \\ k \in \{1, ..., K^c\}\}, \quad (1)$$

where $\text{dom}(f_k^c)$ represents the domain of function f_k^c, which is the set of points in $\mathbb{R}^2$ on which f_k^c is defined, e.g., the coordinates of points in Figure 5, the term $\text{diam}(\text{dom}(f_k^c))$ refers to the diameter of the set $\text{dom}(f_k^c)$, e.g., the diameter of point clouds in Figure 5, and $\sigma(\text{diam}(\text{dom}(f_k^c)))$ is the the visual angle Θ^c spanned by that diameter. Similarly, ψ^c is the visual angle spanned by the diameter of a central subset of $\text{dom}(f_k^c)$ called the fovea. The parameter $\mathcal{D}^c$ is used to specify a two dimensional spatial distribution of points in $\text{dom}(f_k^c)$.



(a) (b)

Fig. 5: Two example point clouds representing visual layers according to the definition in (1). A different distribution $\mathcal{D}^c$ is used for each cloud. In (a), the distribution $\mathcal{D}^c$ is chosen as a regular grid. This distribution is more adapted to representing images with a classical rectangular shape. In (b), this distribution is chosen at random. This shows that the representation of visual layers in the proposed framework is not limited to rectangular distributions as in most vision models.

As an illustrative example, the definition in (1) can be used to represent a classical two dimensional RGB image. In this case, the distribution $\mathcal{D}^c$ is chosen as a 2D rectangular grid corresponding to pixel positions of the image as in Figure 5 (a), k refers to an image component (R, G or B) and f_k^c is the value of the component k at an index (i, j) in $\mathbb{N}^2$.

An interesting feature of using $\mathcal{C}$ to represent an image is that it associates a visual angle Θ^c with its diagonal. This emulates the fact that, in reality, an image is always associated with a certain visual angle when viewed from a certain distance. We argue that this is an important element for any model

that aims at a faithful modeling of the visual systems. It allows to study the influence of the visual angle on the behavior of models performing visual tasks.

The cone receptors layer in the ONL can also be modeled using the definition in (1). In this case, $\mathcal{D}^c$ would be chosen to approximate cone distribution in the retina shown in Figure 3. This means that the density of points in $\mathrm{dom}(f_k^c)$ would be higher in the foveal region defined by ψ^c, and decrease logarithmically towards the periphery. Each point f_k^c would represent a cone receptor whose type would be determined by the subscript k (a S, M or L cone). The angles Θ^c and ψ^c would represent the layer's visual field and the width of the fovea in degrees of visual angles, respectively.

In a much similar way to representing a cone-receptors layer, other visual layers such as the ganglion cell layer in the retina, LGN, V1 and higher layers can be modeled by (1) as we will see in Section 5.

An optional modulation function m^c can be applied to a layer $\mathcal{C}$:

$$m^c : \mathbb{R}^{|\mathcal{C}|} \rightarrow \mathbb{R}^{|\mathcal{C}|}. \tag{2}$$

This function can be used to implement any operation that globally modifies values of points in a given layer. Example operations include non-linearities such as contrast gain control and intensity adaptation as in the retina, the Inhibition of Return (IOR) operation used in most models of visual attention, or any other operation.

4.3 Stacking layers

In the same way as a ganglion cell pools over a number of photo-receptors (mediated by bipolar cells), or a neuron in V1 pools over a number of axons seen by its RF in LGN to produce their output, layers of type $\mathcal{C}$ can be stacked to emulate the feed-forward path of the visual system. In this case, each point in a $\mathcal{C}$-type layer gets its value by pooling over a set of points belonging to the previous layer. More precisely, given two layers $\mathcal{C}_1$ and $\mathcal{C}_2$, a point $f_k^{c_2}(x_o, y_o) \in \mathcal{C}_2$ can be associated with a set of coordinates called its receptive field $\mathrm{RF}^{c_1 c_2}(f_k^{c_2}) \subseteq \mathrm{dom}(f_k^{c_1})$, where $f_k^{c_1} \in \mathcal{C}_1$:

$$\begin{aligned} \mathrm{RF}_k^{c_1 c_2}(f_k^{c_2}(x_o, y_o)) = \{(x, y) | (x, y) &\in \mathrm{dom}(f_{k'}^{c_1}), \\ (x_o, y_o) &\in \mathrm{dom}(f_k^{c_2}), \\ &\text{and } (x, y) \text{ satisfies some condition} \\ &\text{guaranteeing its membership to the} \\ &\text{receptive field of } f_k^{c_2}(x_o, y_o)\}. \end{aligned} \tag{3}$$

Determining whether a coordinate (x, y) is in the receptive field of a point $f_k^{c_2}(x_o, y_o)$ depends on the types of $\mathcal{C}_1$ and $\mathcal{C}_2$. For example, if both $\mathcal{C}_1$ and $\mathcal{C}_2$ model cortical layers, then a typical way of determining the RF membership is by looking whether (x, y) falls within a disk-shaped region around (x_o, y_o),

given that (x, y) and (x_o, y_o) belong to the same space. When $\mathcal{C}_1$ is used to model a RGB image, and $\mathcal{C}_2$ models a cone-receptor layer, the process becomes similar to retinal sampling where a point in the receptors layer gets its value by sampling only one pixel in the image. In this case, determining the receptive field of $f_k^{c_2}(x_o, y_o)$ consists in finding its corresponding point in $\mathcal{C}_1$.

The input signal to the point $f_k^{c_2}(x_o, y_o)$ can be defined as follows:

$$s_k^{c_1 c_2}(f_k^{c_2}(x_o, y_o)) = \{f_{k'}^{c_1}(x, y) | \\ (x, y) \in \mathrm{RF}^{c_1 c_2}(f_k^{c_2}(x_o, y_o))\}, \tag{4}$$

and the value of $f_k^{c_2}(x_o, y_o)$ can be finally computed as:

$$f_k^{c_2}(x_o, y_o) = \phi_k^{c_1 c_2}(s_k^{c_1 c_2}(f_k^{c_2}(x_o, y_o))), \tag{5}$$

where $\phi^{c_1 c_2}$ is a mapping defined as:

$$\phi_k^{c_1 c_2} : \mathbb{R}^{|s_k^{c_1 c_2}(f_k^{c_2}(x_o, y_o))|} \to \mathbb{R}. \tag{6}$$

This mapping can be linear as in the case of Gabor or DoG kernels. It can also be used to implement non-linearities for pooling functions.

In the next section, we propose a model for saliency-based visual attention that implements the proposed vision framework. This will shed the light on the framework's interesting properties and raises some insightful questions about their role in visual processing in Section 6.

5 Application: modeling bottom-up visual attention

Many models have been proposed in litterature for modeling visual attention in recent years. This emerging field has been the subject of a large body of research in neuroscience as well as in computer vision. It has been useful in many applications including object recognition and video compression (Borji and Itti, 2013; Walther et al, 2004), object segmentation (Tu et al, 2016) and detection (Pan et al, 2016; Gao et al, 2015).

The Feature Integration Theory (FIT) introduced in (Treisman and Gelade, 1980) was probably the first to suggest a fundamental functional role for attention in visual recognition. A few years later, Koch & Ullman proposed a possible neural mechanism for driving attention (Koch and Ullman, 1987). This mechanism only considered low-level image features in which only color, intensity contrast and local intensity orientations are used to drive the focus of attention. The first working implementation of this mechanism was proposed by Itti and Koch, and became a landmark for saliency prediction based on bottom-up visual attention (Itti et al, 1998). The term bottom-up comes from the fact that only basic information about the image signal such as color and intensity are involved in predicting saliency. Other models has attempted to

enforce bottom-up biases with higher level information about the scene such as recognition of objects or proto-objects (Judd et al, 2009; Zhao et al, 2014), scene context and gist information (Goferman et al, 2012; Torralba et al, 2006) and by using fully convolutional neural networks more recently (Kruthiventi et al, 2015).

The algorithm we propose here is based on the model of Itti and Koch. However, it shortcuts the first two steps consisting in Gaussian sub-sampling and across-scale subtraction. These steps are replaced by a filtering operation using kernels with eccentricity-dependent receptive fields emulating the cortical magnification effect. This allows us to reduce the number of feature maps to 9 maps instead of 42 maps in the original model.

The model we propose holds some similarity to the one the authors introduced in (Aboudib et al, 2015) with several major differences:

- The proposed model is implemented using the vision framework proposed in Section 4.
- The proposed vision framework allows for a more plausible way for emulating retinal sampling and cortical magnification factors.
- Normalized feature maps are directly combined to form the final saliency map without computing conspicuity maps.

Figure 6 depicts the basic architecture of the attention algorithm based on the proposed vision framework.

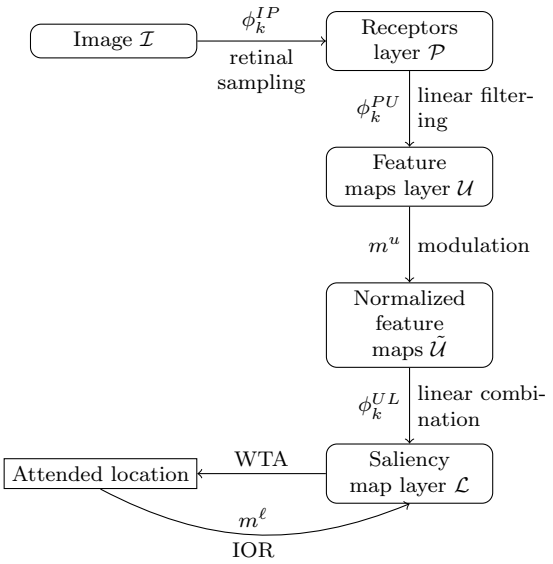


Fig. 6: The basic architecture of the proposed attention algorithm.

5.1 The image layer $\mathcal{I}$

The attention model we propose consists of four $\mathcal{C}$-type layers called $\mathcal{I}$, $\mathcal{P}$, $\mathcal{U}$ and $\mathcal{L}$ defined according to (1). The first layer $\mathcal{I}$ represents an RGB image and is defined as follows:

$$\begin{aligned}\mathcal{I}(\Theta^I, \psi^I, \mathcal{D}^I) = \{f_k^I | f_k^I : \mathbb{N}^2 \to [0,1], \\ \sigma(\text{diam}(\text{dom}(f_k^I))) = \Theta^I, \\ k \in \{1, 2, K^I = 3\}\},\end{aligned} \tag{7}$$

where $\text{dom}(f_k^I)$ is the set of all pixel indexes (i, j) in the image. A point f_k^I represents the value of the k component of the RGB image $\mathcal{I}$ at a given index in $\mathbb{N}^2$, where $k = 1$ stands for the R component, $k = 2$ for G and $k = 3$ for the blue component B.

5.2 The receptors layer $\mathcal{P}$

The second layer $\mathcal{P}$ is the receptors layer that samples the input image in the same way the ONL layer in the retina samples the visual scene, defined similarly as:

$$\begin{aligned}\mathcal{P}(\Theta^p, \psi^p, \mathcal{D}^p) = \{f_k^p | f_k^p : \mathbb{R}^2 \to [0,1], \\ \sigma(\text{diam}(\text{dom}(f_k^p))) = \Theta^p, \\ k \in \{1, 2, K^p = 3\}\},\end{aligned} \tag{8}$$

where the distribution $\mathcal{D}^p$ is chosen to approximate the cone distribution in the primate retina as in Figure 7. A point f_k^p represents a cone receptor of type L or red ($k = 1$), M or green ($k = 2$), S or blue ($k = 3$).

Point coordinates (r, w) in $\text{dom}(f_k^p)$ are expressed in degrees, where r is the eccentricity relative to the center of the fovea measured in degrees of visual angles, which is a typical way of referring to cell positions in the retina. The coordinate w is the angle made between the horizontal line passing through the fovea's center and the line between the fovea's center and (r, w). Parameters ψ^p and Θ^p are also expressed in visual angles. They refer to the diameter span of the fovea and the overall visual field of $\mathcal{P}$, respectively. Figure 7 depicts an example distribution of points in $\text{dom}(f_k^p)$. Notice that points are very dense toward the center where the fovea is found and get sparser toward the periphery.

In order to compute the value of a point $f_k^p \in \mathcal{P}$, a mapping ϕ_k^{Ip} is applied. This mapping can be viewed as a retinal sampling operation where each point in $\mathcal{P}$ is used to sample only one pixel of the image $\mathcal{I}$ at the corresponding location. Hence, given the distribution $\mathcal{D}^p$, the image is sampled at the highest resolution in the fovea, and at progressively lower resolutions toward the periphery.

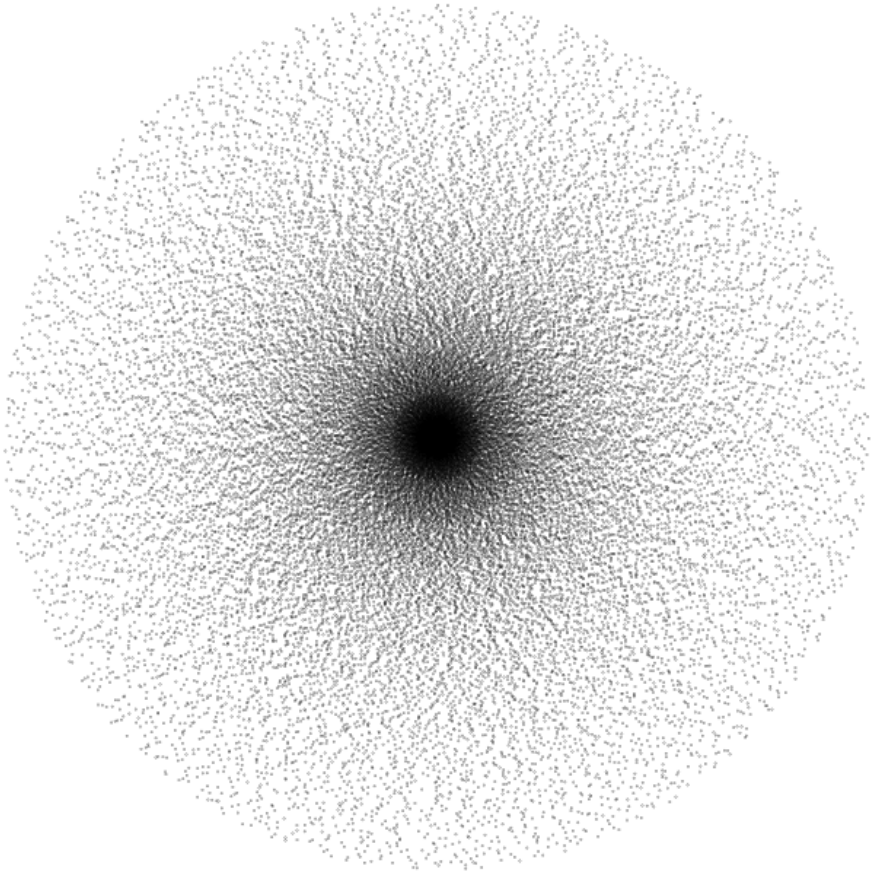


Fig. 7: The distribution $\mathcal{D}^p$ used for the receptor layer $\mathcal{P}$. This distribution is inspired by the distribution of cone receptors in the retina, where the density is higher in the central fovea and decreases rapidly toward the periphery. In this figure, the span of the diameter of layer $\mathcal{P}$ is $\sigma(\text{diam}(\text{dom}(f_k^p))) = \Theta^p = 10^\circ$ and the span of the fovea diameter $\psi^p = 1^\circ$. The total number of points in this figure is 41284 of which 10000 are within the fovea.

We start by determining the set $\text{RF}_k^{Ip}(f_k^p(r_o, \omega_o))$ as:

$$\begin{aligned} \text{RF}_k^{Ip}(f_k^p(r_o, \omega_o)) = \{(i,j) | &(i,j) \in \text{dom}(f_{k'}^I), \\ &(r_o, \omega_o) \in \text{dom}(f_k^p), \\ &\text{and } (i,j) = \text{proj}^{pI}(r_o, \omega_o)\}, \end{aligned} \tag{9}$$

where proj^{pI} is a mapping that associates with each coordinate (r_o, ω_o) in $\text{dom}(f_k^p)$ an index (i,j) in $\text{dom}(f_k^I)$:

$$\text{proj}^{pI} : \mathbb{R}^2 \rightarrow \mathbb{N}^2. \tag{10}$$

Since the radial coordinate r_o is expressed in degrees of visual angles, a natural graphical representation of layer $\mathcal{P}$ would be a spherical surface as in Figure 8. This is close to the real shape of the primate retina, which is often modeled by a spherical surface centered around the nodal point of the eyeball. This spherical representation is used for all subsequent layers of the proposed model: $\mathcal{U}$ and $\mathcal{L}$. Given this representation, the mapping proj^{pI} can be determined from Figure 8 as follows:

$$\begin{aligned} (r, \omega) &= (d \tan(r_o), \omega_o), \\ (i, j) &= \text{proj}(r_o, \omega_o) = ([x], [y]), \end{aligned} \tag{11}$$

where [.] denotes a rounding operation to the closest integer value, and d emulates the distance between the photo plane and the nodal point of the eyeball as shown in Figure 8:

$$d = \frac{\text{diam}(\text{dom}(f_k^I))}{\tan \Theta^I}. \tag{12}$$

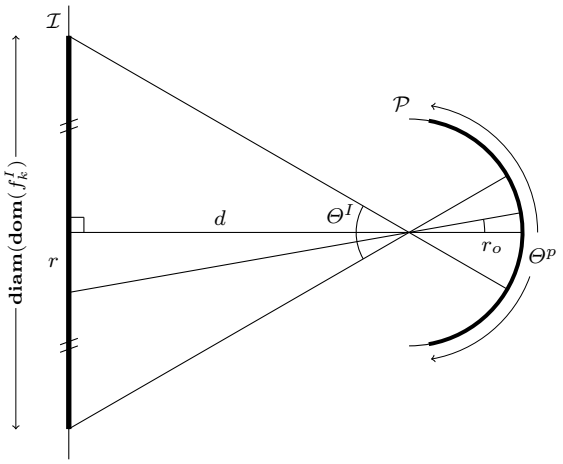
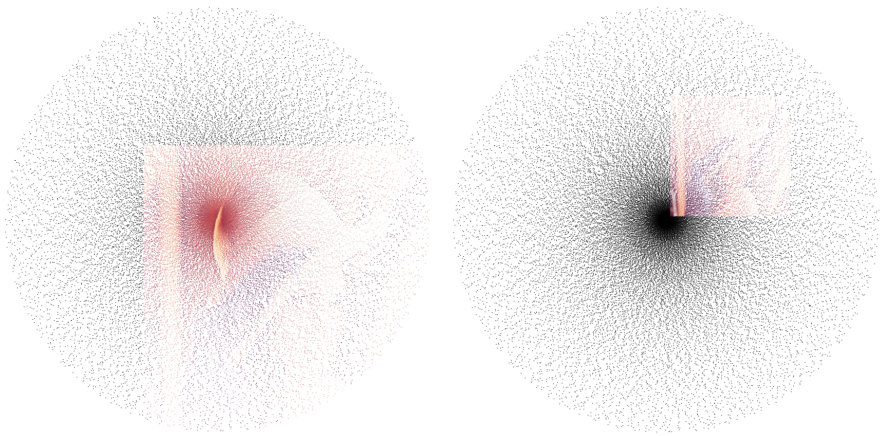


Fig. 8: The projection of the image layer $\mathcal{I}$ onto the receptor layer $\mathcal{P}$ modeled by a hemisphere. The figure is a 2-dimensional cross section plane that passes through the diagonal of the image $\mathcal{I}$ and through the center of the sphere. The point of the image whose radial coordinate is r falls onto the image diagonal.

Notice from (11) that setting the value of ω to ω_o ignores the fact the the image is inverted on the surface of layer $\mathcal{P}$ as shown is Figure 8. A more

faithful way would be to set ω to $\omega_o + \pi$. However, this inversion can be safely ignored since it has no significance on the visual processing task in question.

Also notice that in our experiments, we always consider that the center of the fovea is fixated at the image's center as in Figure 10. However, this model offers the possibility to fixate the fovea at any arbitrary point of the image or even outside its borders as shown in Figure 9, which is a useful property for designing models that needs to emulate saccadic eye movements.



(a) (b)

Fig. 9: The acquisition of the signal of layer $\mathcal{P}$ is totally independent of the size, position and the resolution of the image $\mathcal{I}$. In (a), the visual angle of the image is set to $\Theta^I = 10°$, but the fovea falls onto the upper-left corner of the image so that a part of the image falls outside of the visual field of layer $\mathcal{P}$. In (b), the fovea center falls outside of the image borders, the value of Θ^I is set to $4°$.

The input signal to the point $f_k^p(r_o, \omega_o)$ is then defined as:

$$\begin{aligned} s_k^{Ip}(f_k^p(r_o,\omega_o)) = \{ & f_{k'}^I(i,j) | \\ & (i,j) \in \mathrm{RF}_k^{Ip}(f_k^p(r_o,\omega_o)), \\ & k = k', \\ & \text{and } f_{k'}^I(i,j) \in \mathcal{I}\}, \end{aligned} \tag{13}$$

and finally, sampling is applied by computing the value of each point $f_k^p(r_o, \omega_o)$ as follows:

$$f_k^p(r_o,\omega_o) = \phi_k^{Ip}(s_k^{Ip}(f_k^p(r_o,\omega_o))), \tag{14}$$

where ϕ^{Ip} is a mapping defined on a given set A as follows:

$$\forall A, \phi_k^{Ip}(A) = \begin{cases} A & if A \neq \phi. \\ 0 & \text{Otherwise,} \end{cases} \tag{15}$$

where ϕ is the empty set. Notice that the multi-part definition in (15) accounts for the fact that when $\text{proj}^{pI}(r_o, \omega_o)$ falls outside the image borders, the sampled value is considered as a zero. This is equivalent to considering that the image $\mathcal{I}$ is embedded into a black background. Figure 10 is an example of a retinal representation in $\mathcal{P}$ of an image $\mathcal{I}$ after applying (15).

Fig. 10: The retinal image $\mathcal{P}$ after sampling image $\mathcal{I}$. Notice that the image is sampled at a higher resolution at the foveal center, and that the resolution decreases toward the periphery. No spatial deformations are introduced, and the number of sampled points depends only on the number of points in $\mathcal{P}$ not on the number of pixels in $\mathcal{I}$.

5.3 The feature map layer $\mathcal{U}$

The next layer, is the feature map layer $\mathcal{U}$. This layer is composed of 9 feature maps representing intensity contrast, color opponency and local orientation selectivity, which are the basic three feature dimensions originally used in (Itti et al, 1998):

$$\begin{aligned}\mathcal{U}(\Theta^u, \psi^u, \mathcal{D}^u) = \{f_k^u | f_k^u : \mathbb{R}^2 \rightarrow \mathbb{R}, \\ \sigma(\text{diam}(\text{dom}(f_k^u))) = \Theta^u, \\ k \in \{1, ..., K^u = 9\}\}.\end{aligned} \tag{16}$$

All points in $\mathcal{U}$ that have the same value for k form a single feature map. The 9 feature maps emerging from the above definition, $\{f_{k=1}^u\}$, $\{f_{k=2}^u\}$, $\{f_{k=3}^u\}$,$\{f_{k=4}^u\}$, $\{f_{k=5}^u\}$, $\{f_{k=6}^u\}$, $\{f_{k=7}^u\}$, $\{f_{k=8}^u\}$ and $\{f_{k=9}^u\}$, are chosen to represent intensity contrast, local orientations for 0°, 45°, 90°, 135° and color opponency for red-green, green-red, blue-yellow, yellow-blue, respectively. The distribution $\mathcal{D}^u$ is chosen to be a circular grid as shown in Figure 12. Notice that as in $\mathcal{P}$, the density of points is higher in the fovea and decreases towards the periphery. Also notice that point coordinates in $\text{dom}(f_k^u)$ are expressed in the same units as coordinates in $\text{dom}(f_k^p)$, and they belong to the same space.

Each point f_k^u in $\mathcal{U}$ has its own receptive field in the receptor layer $\mathcal{P}$ spanning a set of coordinates in $\text{dom}(f_k^p)$. Each such RF is defined as follows:

$$\begin{aligned}\text{RF}_k^{pu}(f_k^u(r_o, \omega_o)) = \{(r, \omega) | (r, \omega) \in \text{dom}(f_{k'}^p), \\ (r_o, \omega_o) \in \text{dom}(f_k^u), \\ \text{and } \|(x, y), (x_o, y_o)\| \leq \rho(r_o, \omega_o)\},\end{aligned} \tag{17}$$

where $\|.,.\|$ is the euclidean distance operator, $\rho(r_o, \omega_o)$ is the eccentricity-dependent radius of a circle centered at (r_o, ω_o), and is given by:

$$\rho(r_o, \omega_o) = \begin{cases} \alpha r_o & \text{if } r_o \geq \frac{\psi^u}{2}. \\ \alpha \frac{\psi^u}{2} & \text{otherwise,} \end{cases} \tag{18}$$

where α is the slope associated with the cortical magnification factor (CMF). Notice that (18) reflects the fact that receptive fields of cells within the fovea of a given layer tend to have roughly equal radii. However, these radii begin to increase linearly at the extremities of the fovea toward the periphery, which is behind the cortical magnification effect observed in the primate visual system (Gattass et al, 1981, 1988; Isik et al, 2011).

Notice that a radius $\rho(r_o, \omega_o)$ is measured in degrees of visual angles. Thus, a more precise way to compute the distance between (x, y) and (x_o, y_o) in (17) is to use the great circle distance according to a spherical geometry defined on layer $\mathcal{U}$. However, the spherical surface model of layers $\mathcal{P}, \mathcal{U}$ and $\mathcal{L}$ is supposed

to be locally plane for simplicity, which allows for computing distances as being locally euclidean.

It is worth pointing out that the distribution $\mathcal{D}^u$ can only be determined if the number, sizes and positions of all receptive fields RF_k^{pu} are known. In other words, this distribution is chosen so that a certain overlap is respected between these RFs; the overlap along the radial line p_r, and the overlap p_c between RFs on the same circle. Figure 11 shows and example configuration of receptive fields $\mathrm{RF}_k^{pu}(f_k^u)$ with overlaps $p_r = p_c = 0.5$, and Figure 12 depicts its corresponding distribution $\mathcal{D}^u$.

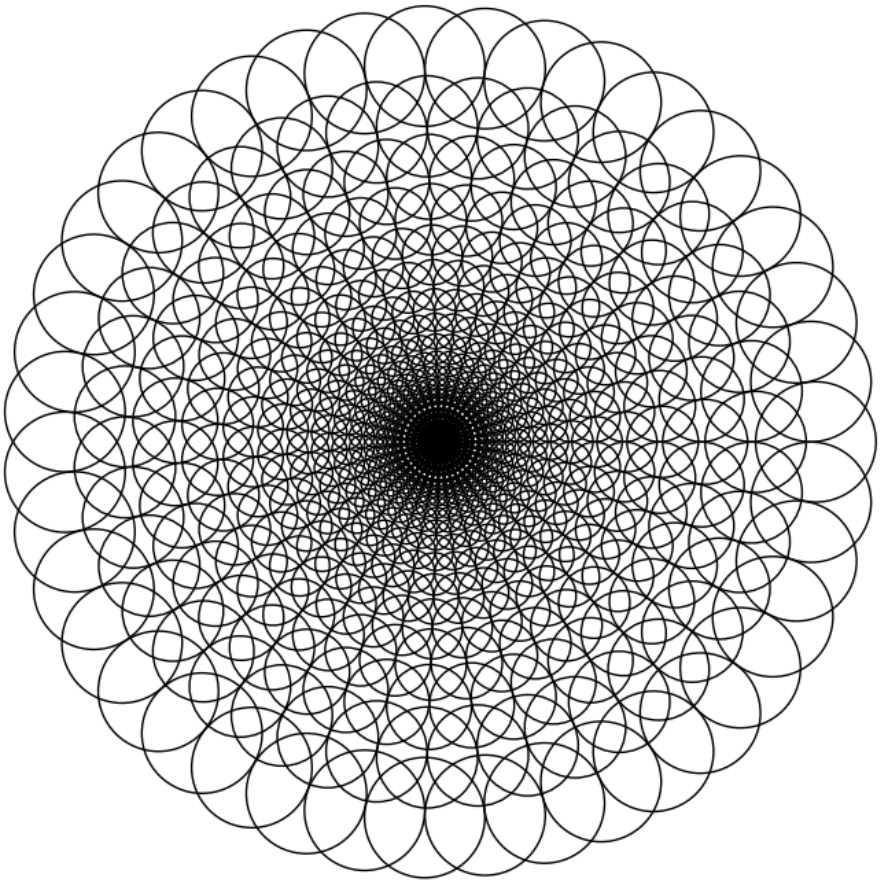


Fig. 11: Receptive fields RF_k^{pu} associated with points in layer $\mathcal{U}$. Notice that these RFs are smaller and more dense in the fovea and grow bigger with decreasing density toward the periphery emulating the cortical magnification factor. This configuration corresponds to circular and radial overlap values $p_c = p_r = 0.5$, a visual angle span of $\Theta^u = 10°$ and a slope $\alpha = 0.16$ for the cortical magnification factor.

The input signals to points belonging to feature maps for intensity contrast and local orientations are given by:

$$s_k^{pu}(f_k^u(r_o,\omega_o))_{k\in\{1,\ldots,5\}} = \{f_{k'}^p(r,\omega)| \\ (r,\omega) \in \mathrm{RF}_k^{pu}(f_k^u(r_o,\omega_o)), \\ k' \in \{1,2,K^p=3\}, \\ \text{and } f_{k'}^p(r,\omega) \in \mathcal{P}\}. \tag{19}$$

Input signals to points within the feature map for red-green opponency are defined as:

$$s_k^{pu}(f_k^u(r_o,\omega_o))_{k=6} = \{f_{k'}^p(r,\omega)| \\ (r,\omega) \in \mathrm{RF}_k^{pu}(f_k^u(r_o,\omega_o)), \\ (k'=1 \wedge \|(x,y),(x_o,y_o)\| \leq (\delta_c/2)) \vee \\ (k'=2 \wedge \|(x,y),(x_o,y_o)\| > (\delta_c/2)), \\ \text{and } f_{k'}^p(r,\omega) \in \mathcal{P}\}, \tag{20}$$

where δ_c is the diameter of the central zone of $\mathrm{RF}_k^{pu}(f_k^u(r'_o,\omega'_o))$ that has a center-surround configuration. We notice from (20) that red-green opponency is applied in the same way as in chromatic ganglion cells that get their input signals from L (red) cones in the central zone of their receptive fields, and from M (green) cones in the surround.

Similarly, input signals for green-red, blue-yellow, yellow-blue feature maps are defined respectively as follows:

$$s_k^{pu}(f_k^u(r_o,\omega_o))_{k=7} = \{f_{k'}^p(r,\omega)| \\ (r,\omega) \in \mathrm{RF}_k^{pu}(f_k^u(r_o,\omega_o)), \\ (k'=2 \wedge \|(x,y),(x_o,y_o)\| \leq (\delta_c/2)) \vee \\ (k'=1 \wedge \|(x,y),(x_o,y_o)\| > (\delta_c/2)), \\ \text{and } f_{k'}^p(r,\omega) \in \mathcal{P}\}, \tag{21}$$

$$s_k^{pu}(f_k^u(r_o,\omega_o))_{k=8} = \{f_{k'}^p(r,\omega)| \\ (r,\omega) \in \mathrm{RF}_k^{pu}(f_k^u(r_o,\omega_o)), \\ (k'=3 \wedge \|(x,y),(x_o,y_o)\| \leq (\delta_c/2)) \vee \\ (k' \in \{1,2\} \wedge \|(x,y),(x_o,y_o)\| > (\delta_c/2)), \\ \text{and } f_{k'}^p(r,\omega) \in \mathcal{P}\}, \tag{22}$$

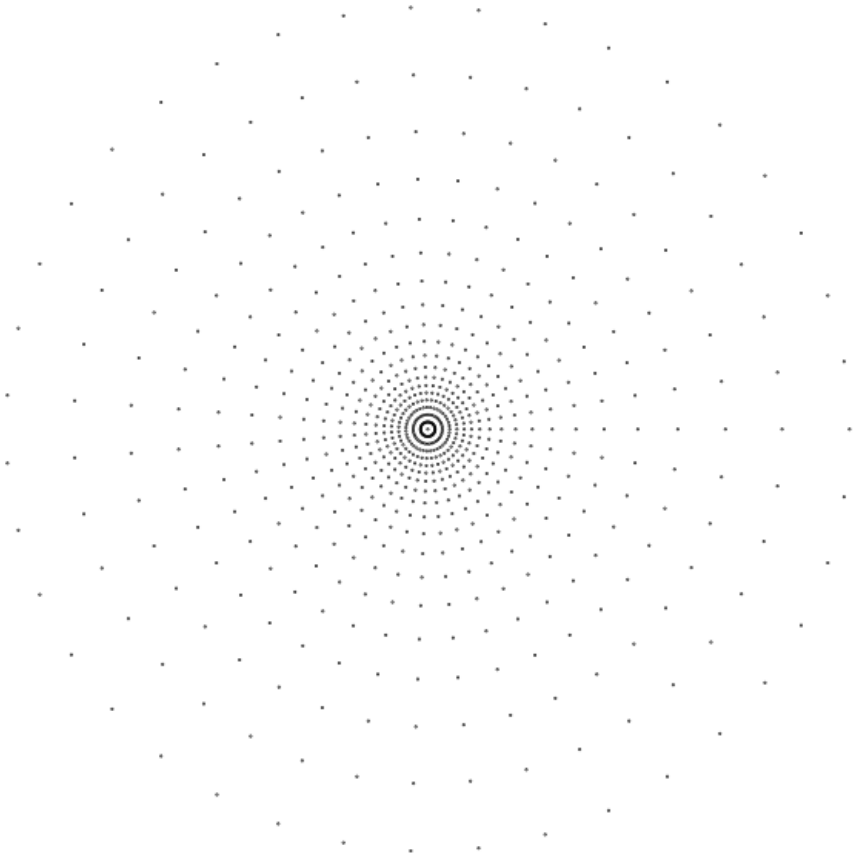


Fig. 12: An example distribution $\mathcal{D}^u$ for points in layer $\mathcal{U}$ corresponding to circular and radial overlap values $p_c = p_r = 0.5$ between the receptive fields RF_k^{pu} associated with each point. This value is chosen for the clarity of display. A value of 0.8 is used for the experiments. The visual angle span of the layer's diameter is $\sigma(\mathrm{diam}(\mathrm{dom}(f_k^u))) = \Theta^u = 10°$.

$$
\begin{aligned}
s_k^{pu}(f_k^u(r_o,\omega_o))_{k=9} = \{ & f_{k'}^p(r,\omega) | \\
& (r,\omega) \in \mathrm{RF}_k^{pu}(f_k^u(r_o,\omega_o)), \\
& (k' \in \{1,2\} \wedge \|(x,y),(x_o,y_o)\| \leq (\delta_c/2)) \vee \\
& (k' = 3 \wedge \|(x,y),(x_o,y_o)\| > (\delta_c/2)), \\
& \text{and } f_{k'}^p(r,\omega) \in \mathcal{P}\}.
\end{aligned} \tag{23}
$$

The value of each point f_k^u in the feature maps is then computed by applying a linear mapping ϕ_k^{pu}.

$$f_k^u(r_o, \omega_o) = \phi_k^{pu}(s_k^{pu}(f_k^u(r_o, \omega_o))). \tag{24}$$

This mapping consists in applying a DoG kernel on each input signal for the intensity contrast and color opponency feature maps, and a Gabor (GB) kernel in feature maps for local orientations. DoG kernels are classically used to model the center-surround configuration of RFs of parasol and midget ganglion cells involved in chromatic and achromatic vision, while GB kernels are typically used to model orientation selective responses of neurons in V1, as mentioned in Section 3.

The DoG model proposed by Rodieck in (Rodieck, 1965) is used to compute the kernel coefficients associated with a point at a coordinate (r, ω).

$$\begin{aligned}\mathrm{DoG}(r_o, \omega_o, r, \omega) =& g_1 \frac{\pi}{\delta_1^2}.\exp(-\frac{\|(x,y),(x_o,y_o)\|^2}{\delta_1^2}) - \\ & g_2 \frac{\pi}{\delta_2^2}.\exp(-\frac{\|(x,y),(x_o,y_o)\|^2}{\delta_2^2}),\end{aligned} \tag{25}$$

where (r_o, ω_o) is the RF center to which (r, ω) belongs, δ_1 and δ_2 are the standard deviations of the central and the surround Gaussians of DoG kernels, g_1 and g_2 are two constants used to control the relative strengths of the two Gaussians.

Coefficients of Gabor kernels (Gabor, 1946) are similarly defined as follows:

$$\mathrm{GB}(r_o, \omega_o, r, \omega) = \exp(-\frac{X^2 + Y^2\gamma^2}{2\delta_3}).\cos(\frac{2\pi}{\lambda}X), \text{ s.t.} \tag{26}$$

$$X = (x - x_o)\cos\theta + (y - y_o)\sin\theta \text{ and}$$

$$Y = -(x - x_o)\sin\theta + (y - y_o)\cos\theta. \tag{27}$$

Figure 13 depicts some examples of DoG and GB kernels we used. The mapping ϕ_k^{pu} is finally applied as the sum of elements of an input signal weighted by their corresponding kernel coefficients:

$$\begin{aligned}&\phi_k^{pu}(s_k^{pu}(f_k^u(r_o, \omega_o)))_{k\in\{1,6,7,8,9\}} = \\ &= \sum_{\substack{f^p(r,\omega)\subseteq \\ s_k^{pu}(f_k^u(r_o,\omega_o))}} \mathrm{mean}(f^p(r, \omega)).\mathrm{DoG}(r_o, \omega_o, r, \omega),\end{aligned} \tag{28}$$

$$\begin{aligned}&\phi_k^{pu}(s_k^{pu}(f_k^u(r_o, \omega_o)))_{k\in\{2,3,4,5\}} = \\ &= \sum_{\substack{f^p(r,\omega)\subseteq \\ s_k^{pu}(f_k^u(r_o,\omega_o))}} \mathrm{mean}(f^p(r, \omega)).\mathrm{GB}(r_o, \omega_o, r, \omega),\end{aligned} \tag{29}$$

where $f^p(r, \omega)$ here is the set of all points $f_{k'}^p(r, \omega)$ in $s_k^{pu}(f_k^u(r_o, \omega_o))$ defined on the same coordinate (r, ω). Figure 14 is an example of some feature maps we obtain by applying the above mappings.

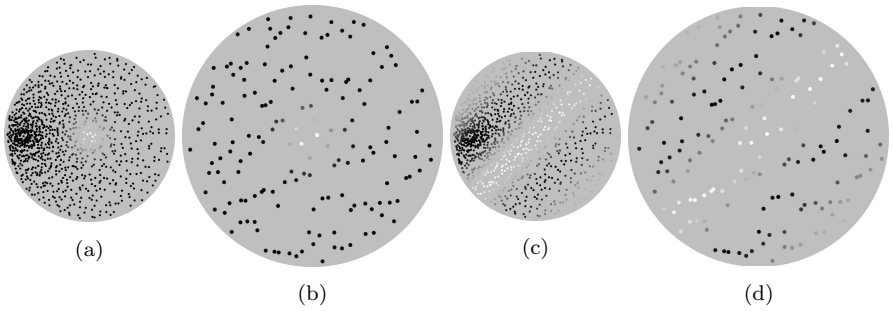


Fig. 13: Some examples of the Difference of Gaussians (DoG) and Gabor (GB) kernels used for the mapping ϕ_k^{pu}. Notice that kernels whose RFs are closer to the fovea in (a) and (c) are smaller in size in and defined on more points that RFs in the periphery, (b) and (d), which is due to the cortical magnification factor. This difference in size and density is inspired by biological reality in the retina. The gray background and the size of points in these figures is adjusted for the clarity of display.

5.4 The saliency map layer $\mathcal{L}$

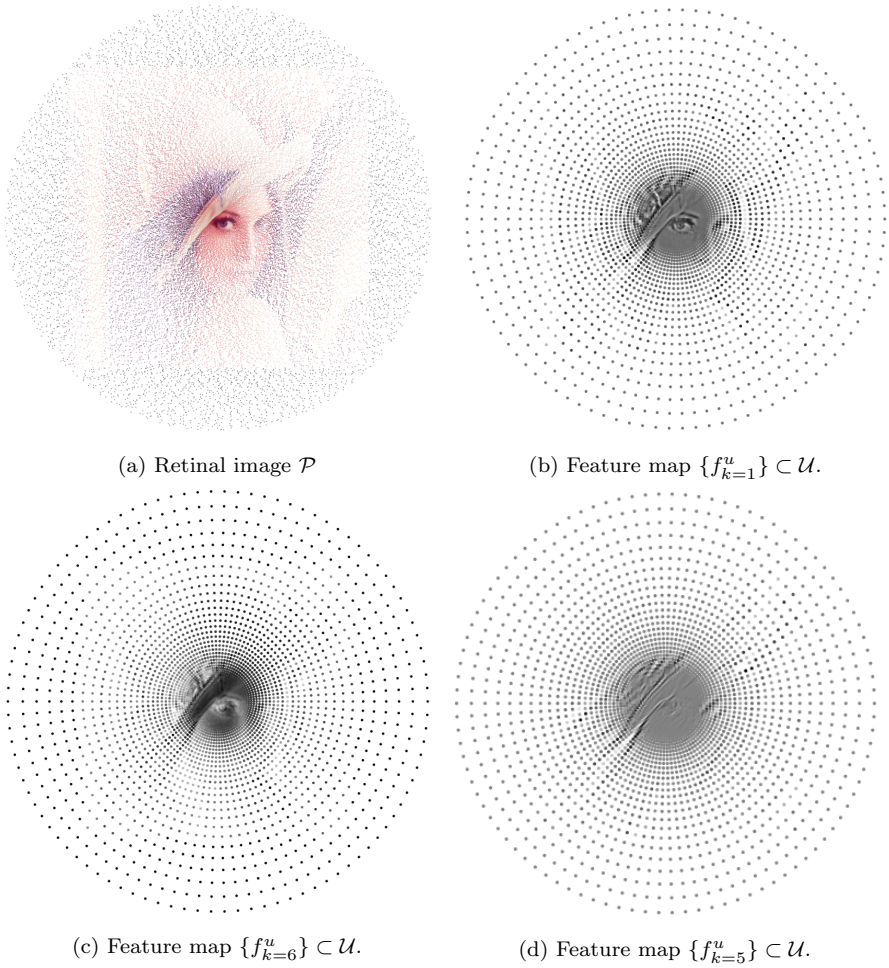
Finally, layer $\mathcal{L}$ is used to compute the saliency map:

$$\begin{aligned}\mathcal{L}(\Theta^\ell, \psi^\ell, \mathcal{D}^\ell) = \{f_k^\ell | f_k^\ell : \mathbb{R}^2 \rightarrow \mathbb{R},\\ \sigma(\text{diam}(\text{dom}(f_k^\ell))) = \Theta^\ell,\\ \mathcal{D}^\ell = \mathcal{D}^u,\\ \text{and } k \in \{K^\ell = 1\}\},\end{aligned} \tag{30}$$

This saliency map has exactly the same distribution of point coordinates as that of feature maps. The RF of each point in $\mathcal{L}$ at a coordinate (r_o, ω_o) spans only the point at the same location in $\mathcal{U}$.

$$\begin{aligned}\text{RF}_k^{u\ell}(f_k^\ell(r_o, \omega_o)) = \{(r, \omega) | (r, \omega) \in \text{dom}(f_k^u),\\ (r_o, \omega_o) \in \text{dom}(f_k^\ell),\\ \text{and } (r, \omega) = (r_o, \omega_o)\}.\end{aligned} \tag{31}$$

Before point values in $\mathcal{L}$ could be computed, a modulation function m^u as defined in (2) is applied to $\mathcal{U}$. This modulation is used to increase the contrast of the most salient regions in each feature map in a similar way the map normalization operator $\mathcal{N}(.)$ is applied in (Itti et al, 1998).



(a) Retinal image $\mathcal{P}$

(b) Feature map $\{f^u_{k=1}\} \subset \mathcal{U}$.

(c) Feature map $\{f^u_{k=6}\} \subset \mathcal{U}$.

(d) Feature map $\{f^u_{k=5}\} \subset \mathcal{U}$.

Fig. 14: The retinal image represented by the receptors layer $\mathcal{P}$ (a) and some corresponding feature maps held by layer $\mathcal{U}$: intensity contrast feature map (b), red-green opponency feature map (c) and 135°-orientation feature map (d). These feature maps correspond to circular and radial overlap values $p_c = p_r = 0.8$, a visual angle span of $\Theta^I = 10°$ and a slope $\alpha = 0.16$ for the cortical magnification factor.

$$\begin{aligned}\tilde{\mathcal{U}} = m^u(\mathcal{U}) = \{f^{\tilde{u}}_k | f^{\tilde{u}}_k : \mathbb{R}^2 \rightarrow [0,1], \\ \mathrm{dom}(f^{\tilde{u}}_k) = \mathrm{dom}(f^u_k), \\ k \in \{1, ..., k^u = 9\}\}.\end{aligned} \tag{32}$$

The steps for computing the value of the modulated points $f^{\tilde{u}}_k$ are the following:

1. A half-wave rectification is first applied to feature maps to remove negative values.

$$f_k^{\tilde{u}} = \max(0, f_k^u). \tag{33}$$

2. The values within each feature map are scaled to the interval $[0, 1]$.

$$f_k^{\tilde{u}} \leftarrow \frac{f_k^{\tilde{u}} - \min\limits_k(f_k^{\tilde{u}})}{\max\limits_k(f_k^{\tilde{u}}) - \min\limits_k(f_k^{\tilde{u}})}. \tag{34}$$

3. A multiplicative factor β_k is computed.

$$\beta_k = \left(\max_k(f_k^{\tilde{u}}) - \operatorname*{mean}_k(f_k^{\tilde{u}})\right)^2. \tag{35}$$

4. The multiplicative factor β_k is then applied to each point of the feature maps.

$$f_k^{\tilde{u}} \leftarrow \beta_k f_k^{\tilde{u}} \tag{36}$$

The input signal to each point in the saliency map can now be defined on the modulated feature maps:

$$\begin{aligned} s_k^{u\ell}(f_k^\ell(r_o, \omega_o)) = \{ & f_{k'}^{\tilde{u}}(r, \omega) | \\ & (r, \omega) \in \mathrm{RF}_k^{u\ell}(f_k^\ell(r_o, \omega_o)), \\ & f_{k'}^{\tilde{u}} \in \tilde{\mathcal{U}}. \}. \end{aligned} \tag{37}$$

Finally, the saliency map is computed using the mapping $\phi_k^{u\ell}$, which is the mean of all modulated feature maps in $\tilde{\mathcal{U}}$.

$$\begin{aligned} f_k^\ell(r_o, \omega_o) &= \phi_k^{u\ell}(s_k^{u\ell}(f_k^\ell(r_o, \omega_o))) \\ &= \mathrm{mean}(s_k^{u\ell}(f_k^\ell(r_o, \omega_o))). \end{aligned} \tag{38}$$

5.5 Creating fixation maps

Fixation maps are created by an iterative processes consisting of a Winner-Take-All (WTA) step, which extracts the coordinates of the most salient point in the saliency map followed by an Inhibition-of-Return (IOR) step, which guarantees that previously fixated locations should no longer be visited in subsequent iterations. Here are the details of these two steps:

1. A fixation location (r_o, ω_o) is extracted from the saliency map $\mathcal{L}$.

$$(r_o, \omega_o) = \operatorname*{argmax}_{(r,w)} f_k^\ell. \tag{39}$$

2. IOR is applied using a modulation function m^ℓ.

$$m_k^\ell(\mathcal{L}) :$$
$$f_k^\ell(r,\omega) \leftarrow \begin{cases} f_k^\ell(r,\omega) \text{ if } \|(x,y),(x_o,y_o)\| > h \\ 0 \qquad\qquad\quad \text{otherwise,} \end{cases} \tag{40}$$

   where h is the radius of the inhibited zone in visual angles.
3. the pixel indexes (i,j) in the image $\mathcal{I}$ corresponding to the fixation location (r_o,ω_o) are then computed using (11).

Figure 15 depicts an example of a saliency map in layer $\mathcal{L}$ and the corresponding smoothed saliency map. A smoothed saliency map is one consisting in convoluting a gaussian kernel on the extracted fixation locations, in order to produce a continuous gray-scale saliency map of the same size as the input image $\mathcal{I}$.

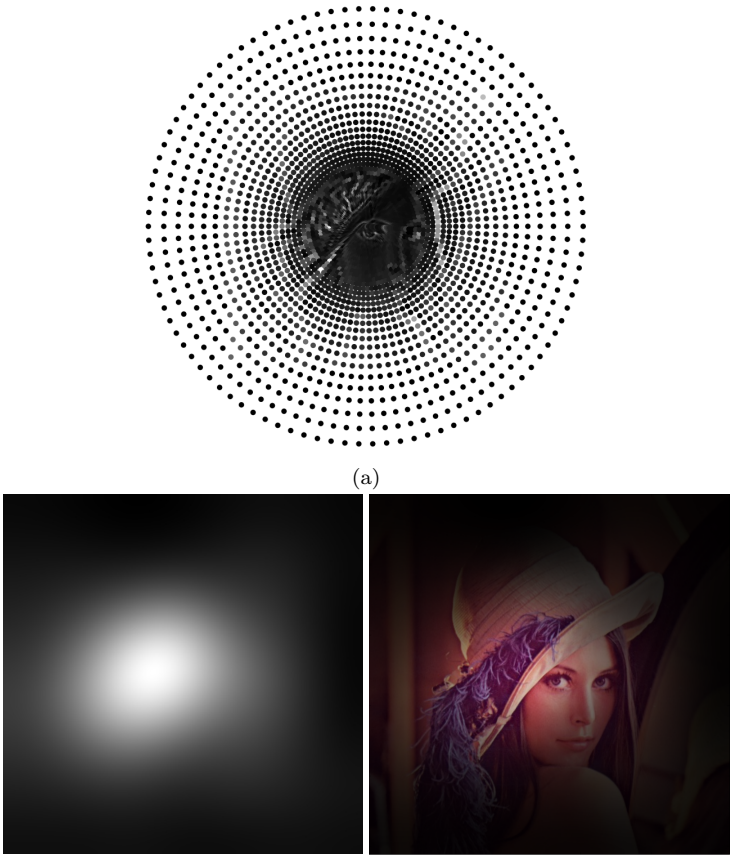
In the next section, we provide a performance evaluation of the proposed attention model along with a comparison with some of the state-of-the-art models, and a discussion of the results.

6 Results and discussion

In order to validate the performance of the proposed model on estimating bottom-up visual saliency, we ran the algorithm on the CAT2000 test dataset provided by the MIT saliency benchmark. This dataset contains 2000 images from 20 different categories with a fixed size of 1920×1080 pixels (Borji et al, 2013a).

Before beginning the test on the above dataset, we performed a minor optimization of the model parameters on the CAT2000 train dataset containing 2000 images of the same 20 categories as in the CAT2000 test set (Borji et al, 2013a).

Hence, we set the model parameters as follows: The visual angles $\Theta^I = \Theta^p = \Theta^u = \Theta^\ell = 10°$, which represent the visual field available to the system. The diameter of the fovea of all layers $\psi^I = \psi^p = \psi^u = \psi^\ell = 1°$. The total number of receptors in $\mathcal{P}$ is set to 123853 of which 30000 are within the fovea. This is equivalent to 41284 total RGB pixels of which 10000 are within the fovea, as shown in Figures 7 and 10. This means that retinal images used by the algorithm has more than 50 times less pixels than the original images, which represents a significant reduction of information. The overlap parameters p_r and p_c between RFs of points in $\mathcal{U}$ are both set to 0.8. The slope α in (18) associated with the cortical magnification factor in layer $\mathcal{U}$ is set to 0.16 which is close to its value in layer $V1$ of the ventral stream found by Gattass in (Gattass et al, 1988). For each image, the 250 most salient fixations locations are extracted by an iterative WTA and IOR process. The radius h of the inhibited zone at each IOR iteration is set to 0.05° of visual angles.



(a) (b) (c)

Fig. 15: An example of a saliency map carried by layer $\mathcal{L}$ (a) where the size of single points is adapted for a better clarity of display. The corresponding smoothed saliency map is shown in (b) made by convoluting a Gaussian kernel on the first 250 fixation locations. In (c), the smoothed saliency map is superimposed on the original image $\mathcal{I}$.

Parameters of DoG kernels were adapted from (Rodieck, 1965). We set g_2/g_1 to 0.8, δ_2/δ_1 to 3 and ρ/δ_1 to 11.8, where g_2/g_1 is a measure of the ratio of strength of the surround to the center Gaussians of the DoG kernel, and δ_1 and δ_2 are the effective widths of the center and surround Gaussians, respectively.

For Gabor kernels, we adapted parameter values used for designing simple cells in the Hmax model (Serre et al, 2007); The aspect ratio is set as $\gamma = 0.3$. We also set $\delta_3/\lambda = 0.8$ and $\rho/\delta_3 = 2.5$, where δ_3 is the effective width of the Gaussian component of the filter, λ is the wavelength of the cosine component,

and ρ in both DoG and Gabor kernels denotes the eccentricity-dependent radius of a given kernel computed from (18) and expressed in visual angles.

Figure 16 depicts some examples of images taken from the CAT2000 test dataset and their corresponding smoothed saliency maps.

Fig. 16: Some example images from the CAT2000 test dataset (left column), the corresponding smoothed saliency maps (middle column) and with saliency maps superimposed (right column).

Table 1 shows the scores of our models according to several metrics used by the benchmark and how they compare to other models. This table and more detailed comparisons are also available on the MIT Saliency Benchmark website `http://saliency.mit.edu/results_cat2000.html`.

As shown in Table 1, the proposed model shows good performance scores relative to other models. These scores are computed according to 7 metrics: the Similarity (Sim), the Correlation-Coefficient (CC), the Normalized Scanpath Saliency (NSS) and Earth Mover's Distance (EMD) and the Area Under the ROC Curve metrics.

It is worth pointing out that the IttiKoch2, GBVS, Judd's and several other models are optimized for smoothing parameters and center-bias. The proposed model, has only a minor optimization for the width of the Gaussian kernel used for smoothing fixation maps while no explicit center-bias is applied. However, such bias arises naturally in the model due to retinal sampling

Model		Sim	AUC-Judd	EMD (Lower is better)	AUC-Borji	CC	NSS	sAUC
Proposed model		**0.58**	**0.80**	**2.10**	**0.77**	**0.64**	**1.57**	**0.55**
BMS	Zhang and Sclaroff (2013)	0.61	0.85	1.95	0.84	0.67	1.67	0.59
GBVS	Harel et al (2006)	0.51	0.80	2.99	0.79	0.50	1.23	0.58
Context-Aware saliency	Goferman et al (2012)	0.50	0.77	3.09	0.76	0.42	1.07	0.60
AWS	Garcia-Diaz et al (2012)	0.49	0.76	3.36	0.75	0.42	1.09	0.62
IttiKock2		0.48	0.77	3.44	0.76	0.42	1.06	0.59
WMAP	López-García et al (2011)	0.47	0.75	3.28	0.69	0.38	1.01	0.60
Judd model	Judd et al (2009)	0.46	0.84	3.61	0.84	0.54	1.30	0.56
Torralba saliency	Torralba et al (2006)	0.45	0.72	3.44	0.71	0.33	0.85	0.58
Murray model	Murray et al (2011)	0.43	0.70	3.79	0.70	0.30	0.77	0.59
SUN saliency	Zhang et al (2008)	0.43	0.70	3.42	0.69	0.30	0.77	0.57
IttiKock	Itti et al (1998)	0.34	0.56	4.66	0.53	0.09	0.25	0.52
Achanta	Achanta et al (2009)	0.33	0.57	4.45	0.55	0.11	0.29	0.52

Table 1: A performance comparison between the proposed model and other models on the CAT2000 test dataset of the MIT Saliency Benchmark. These results can be found on the MIT saliency benchmark Web page `http://saliency.mit.edu/results_cat2000.html`.

and cortical magnification factors, which allocate more resources to processing central zones of the image than to peripheral ones. It would be interesting to explore the role of retinal sampling and cortical magnification in influencing center-bias that human subjects manifest when free viewing images.

An important point to discuss is the fact the attentional behavior manifested by eye fixations differs as a function of the viewing distance, (or equivalently the visual angle) between the subject and an image (Borji et al, 2013b). However, attention models in Table 1 do not have a direct way for measuring their performance as a function of the visual angle. This makes interpreting the performance of such models against a given saliency dataset more ambiguous and less straightforward. For example, suppose having two datasets D_1 and D_2, associated with two visual angles θ_1 and θ_2, respectively. If a given attention model performs better on D_1 than on D_2, there would be no clear way for determining whether this is due to the fact that it is intrinsically more adapted to the angle θ_1 that to θ_2, or due to other factors.

The model we propose provides the possibility to fix all other parameters while varying the image's visual angle Θ^I. Figures 17 and 18 depict how the performance of the proposed attention algorithm varies according to different evaluation metrics as a function of Θ^I. This provides a mechanism to check whether the model matches ground truth attentional behavior when measured at different visual angles. We think that this is a useful factor to consider for

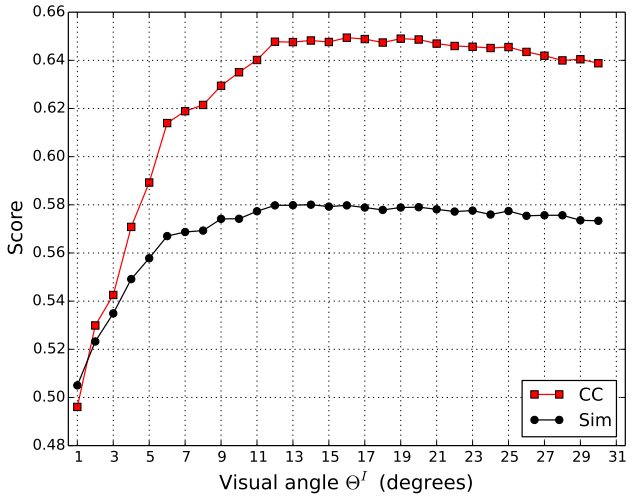


Fig. 17: The influence of changing the images' visual angle Θ^I on the models performance according to the Similarity (Sim) and the Cross Correlation (CC) metrics.

models that seek biological plausibility. However, as to our knowledge, no available benchmarks provide fixation data measured at different visual angles yet. Creating such a benchmark would provide the possibility to analyze and validate our performance curves in Figures 17 and 18, as well as those of future models that might choose to integrate a visual angle parameter.

Finally, while the proposed model does not always give the best saliency prediction according to Table 1, it provides some advantages over other models from a biological point of view:

- Eye movements and fixations can be emulated more faithfully using the proposed vision framework. As in the real retina, moving the fovea over the image will change the resolution perceived at each region of the image due to retinal sampling and cortical magnification factors.
- A more straight-forward way to compare to ground truth on vision tasks. Effects of fundamental vision parameters absent from most saliency models, such as viewer distance, visual field, cortical magnification and retinal sampling could potentially be studied more closely using the proposed framework.

7 Conclusion and future work

In this paper we proposed a new framework for building visual information processing models. This model is more closely inspired by the architecture

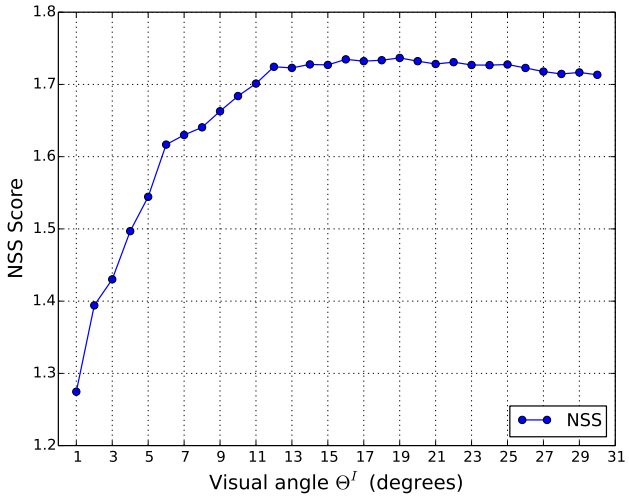


Fig. 18: The influence of changing the images' visual angle Θ^I on the models performance according to the Normalized Scanpath Saliency (NSS) metric.

of the primate visual system, and is motivated by the recent trend in the computer vision community toward a closer modeling of the visual system in the hope of going beyond some limitations in current vision systems.

We have seen that the architecture of the proposed framework offers some interesting properties found in the visual system. For example, the presence of a receptor layer makes the acquired image signal totally independent from the input image's resolution and size. It also motivates the use of such a framework for applications such like saccadic eye movements since the receptor layer is not constrained by the image borders and can be used to receive its signal from any part of the input scene. Moreover, the proposed framework has a very clear notion of a visual angle emulating the ubiquitous presence of this parameter in biological vision. This provides the possibility to better understand the influence of a viewer's distance from an image on vision tasks. Another important property the framework offers is the information reduction by means of retinal sampling and cortical magnification which are two important and omnipresent factors of primate visual systems. We have seen that these two mechanisms can be implemented seamlessly, while avoiding classical problems like spatial deformations and the dependency on the input image size.

In Section 5, we proposed a saliency-driven model of attention built on top of the proposed vision framework. We showed that this model attains state-of-the-art performance. More particularly, we showed that it has a better performance than Itti and Koch's model on which it is based, while using lower resolution and a fewer number of feature maps. This application motivated the use of the proposed vision framework and raises some important

questions such as the role of the visual angle in attention modeling and its importance for a better understanding of attentional behavior and benchmarking results. One possible method we propose to start such exploration, would be to design an attention benchmark that provide eye-fixation data on a given dataset for a range of visual angles. It would be then interesting to study how models' performances should be analyzed and understood given this variability of fixation data associated with different visual angles.

In future work, we will also consider the question of how common architectures for visual processing, especially Convolutional Neural Networks (CNNs) might be adapted for being implemented using the proposed framework. The challenge would be in modifying its learning algorithm so that it can take the cortical magnification factor into account and the associated variability in kernel sizes in each layer.

Another research perspective would be to use the proposed framework for implementing attention-based object recognition processors to account for the retinal transformation stage in models such like (Zheng et al, 2015).

The proposed framework and the associated attention model are already implemented and are publicly available as a Git repository on the Web[1]. However, future work will include further development and improvement of the proposed framework along with its code implementation. We hope that through collaboration, this framework could evolve as an alternative, full-fledged toolbox for neuro-inspired visual processing, in the same way as current programming libraries offer optimized implementations of traditional image processing algorithms.

8 Compliance with Ethical Standards

This study was funded by the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n° 290901. This article does not contain any studies with human participants or animals performed by any of the authors.

References

Aboudib A, Gripon V, Coppin G (2015) A model of bottom-up visual attention using cortical magnification. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pp 1493–1497, DOI 10.1109/ICASSP.2015.7178219

Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, IEEE, pp 1597–1604

Anselmi F, Rosasco L, Poggio T (2015) On invariance and selectivity in representation learning. arXiv preprint arXiv:150305938

[1] https://bitbucket.org/ala_aboudib/see

Bonaiuto J, Itti L (2005) Combining attention and recognition for rapid scene analysis. In: Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, IEEE, pp 90–90

Borji A, Itti L (2013) State-of-the-art in visual attention modeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(1):185–207

Borji A, Sihite DN, Itti L (2013a) Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. Image Processing, IEEE Transactions on 22(1):55–69

Borji A, Tavakoli HR, Sihite DN, Itti L (2013b) Analysis of scores, datasets, and models in visual saliency prediction. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, pp 921–928

Borji A, Sihite DN, Itti L (2014) What/where to look next? modeling top-down visual attention in complex interactive environments. Systems, Man, and Cybernetics: Systems, IEEE Transactions on 44(5):523–538

Dowling JE (1987) The retina: an approachable part of the brain. Harvard University Press

Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nature neuroscience 14(9):1195–1201

Gabor D (1946) Theory of communication. part 1: The analysis of information. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering 93(26):429–441

Gao F, Zhang Y, Wang J, Sun J, Yang E, Hussain A (2015) Visual attention model based vehicle target detection in synthetic aperture radar images: A novel approach. Cognitive Computation 7(4):434–444

Garcia-Diaz A, Leboran V, Fdez-Vidal XR, Pardo XM (2012) On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. Journal of vision 12(6):17–17

Gattass R, Gross C, Sandell J (1981) Visual topography of v2 in the macaque. Journal of Comparative Neurology 201(4):519–539

Gattass R, Sousa A, Gross C (1988) Visuotopic organization and extent of v3 and v4 of the macaque. The Journal of neuroscience 8(6):1831–1845

Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(10):1915–1926

Gonzalez RC, Woods RE (2002) Digital image processing

Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends in neurosciences 15(1):20–25

Harel J, Koch C, Perona P (2006) Graph-based visual saliency. In: Advances in neural information processing systems, pp 545–552

Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology 148(3):574–591

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 160(1):106–154

Isik L, Leibo JZ, Mutch J, Lee SW, Poggio T (2011) A hierarchical model of peripheral vision. Tech. rep., MIT's Computer Science and Artificial Intel-

ligence Laboratory
Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence 20(11):1254–1259
Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, IEEE, pp 2106–2113
Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence, Springer, pp 115–141
Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
Kruthiventi SS, Ayush K, Babu RV (2015) Deepfix: A fully convolutional neural network for predicting human eye fixations. CoRR abs/1510.02927
Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338
Larochelle H, Hinton GE (2010) Learning to combine foveal glimpses with a third-order boltzmann machine. In: Lafferty J, Williams C, Shawe-Taylor J, Zemel R, Culotta A (eds) Advances in Neural Information Processing Systems 23, Curran Associates, Inc., pp 1243–1251
LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324
Lee H, Battle A, Raina R, Ng AY (2006) Efficient sparse coding algorithms. In: Advances in neural information processing systems, pp 801–808
Liu H, Liu Y, Sun F (2015) Robust exemplar extraction using structured sparse coding. IEEE transactions on neural networks and learning systems 26(8):1816–1821
López-García F, Dosil R, Pardo XM, Fdez-Vidal XR (2011) Scene recognition through visual attention and image features: A comparison between sift and surf approaches. INTECH Open Access Publisher
Marčelja S (1980) Mathematical description of the responses of simple cortical cells*. JOSA 70(11):1297–1300
Marr D (1982) Vision, a computational investigation into the human representation and processing of visual information. WH San Francisco: Freeman and Company
Martínez J, Robles LA (2006) A new foveal cartesian geometry approach used for object tracking. SPPRA 6:133–139
McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5(4):115–133
Milner AD, Goodale MA (2008) Two visual systems re-viewed. Neuropsychologia 46(3):774–785
Murray N, Vanrell M, Otazu X, Parraga CA (2011) Saliency estimation using a non-parametric low-level vision model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, IEEE, pp 433–440

Pan J, Li X, Li X, Pang Y (2016) Incrementally detecting moving objects in video with sparsity and connectivity. Cognitive Computation 8(3):420–428

Poggio T, Mutch J, Isik L (2014) Computational role of eccentricity dependent cortical magnification. arXiv preprint arXiv:14061770

Ranzato M, Hinton G, LeCun Y (2015) Guest editorial: Deep learning. International Journal of Computer Vision 113(1):1–2, DOI 10.1007/s11263-015-0813-1

Ray S, Scott S, Blockeel H (2010) Encyclopedia of Machine Learning, Springer US, Boston, MA, chap Multi-Instance Learning, pp 701–710. DOI 10.1007/978-0-387-30164-8_569, URL http://dx.doi.org/10.1007/978-0-387-30164-8_569

Rodieck RW (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vision research 5(12):583–601

Rybak IA, Gusakova V, Golovan A, Podladchikova L, Shevtsova N (1998) A model of attention-guided visual perception and recognition. Vision research 38(15):2387–2400

Salin PA, Bullier J (1995) Corticocortical connections in the visual system: structure and function. Physiological reviews 75(1):107–155

Schwartz EL (1984) Anatomical and physiological correlates of visual computation from striate to infero-temporal cortex. Systems, Man and Cybernetics, IEEE Transactions on (2):257–271

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007) Robust object recognition with cortex-like mechanisms. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(3):411–426

Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review 113(4):766

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cognitive psychology 12(1):97–136

Tu Z, Abel A, Zhang L, Luo B, Hussain A (2016) A new spatio-temporal saliency-based video object segmentation. Cognitive Computation pp 1–19

Walther D, Koch C (2007) Attention in hierarchical models of object recognition. Progress in brain research 165:57–78

Walther D, Rutishauser U, Koch C, Perona P (2004) On the usefulness of attention for object recognition. In: Workshop on Attention and Performance in Computational Vision at ECCV, Citeseer, pp 96–103

Wohrer A, Kornprobst P (2009) Virtual retina: a biological retina model and simulator, with contrast gain control. Journal of computational neuroscience 26(2):219–249

Zhang J, Sclaroff S (2013) Saliency detection: A boolean map approach. In: Proceedings of the IEEE International Conference on Computer Vision, pp 153–160

Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) Sun: A bayesian framework for saliency using natural statistics. Journal of vision 8(7):32

Zhao J, Sun S, Liu X, Sun J, Yang A (2014) A novel biologically inspired visual saliency model. Cognitive Computation 6(4):841–848

Zheng Y, Zemel R, Zhang YJ, Larochelle H (2015) A neural autoregressive approach to attention-based recognition. International Journal of Computer Vision 113(1):67–79

Zhu JY, Wu J, Xu Y, Chang E, Tu Z (2015) Unsupervised object class discovery via saliency-guided multiple class learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on 37(4):862–875