NATURAL SCENE PERCEPTION: VISUAL ATTRACTORS AND IMAGES PROCESSING

A. CHAUVIN, J. HERAULT, C. MARENDAZ AND C. PEYRIN

Laboratoire des Images et des Signaux, CNRS ESA 5083., INPG, 27 Av Félix Viallet, Grenoble 38000 ,
Laboratoire de Psychologie Expérimentale, CNRS UMR 5105., UPMF, Saint Martin-d'Hères 38130 ,
France
E-mail: alan.chauvin; jeanny.herault@lis.inpg.fr & christian.marendaz@upmf-grenoble.fr

This paper aims at identifying the regions of interest in natural scenes. These regions have been defined by a behavioural measure of eye movement and by a model of saliency map constructed in a biologically plausible manner.

The saliency map codes the local region of interest in terms of signal properties such as contrast, orientation, colour, curvature etc. In our approach, pictures are processed using a retinal model, simulating the parvocellular output of the retina. The result is then filtered by a bank of Gabor filters, in mutual interaction in order to lower noise, enhance contour, and sharpen filter selectivity.

Subjects' eye positions were recorded as they explored static black and white images in order to categorize these images. All fixations during one scene were averaged in order to make a density map coding the time spent for subjects on each pixel. Statistics were computed on the regions around the fixation point to evaluate an index of predictability of our saliency map. The saliency map and the density map select similar areas. Furthermore, statistics based on eye-selected regions show greater values than for randomly-selected ones.

1 Introduction

To what extent can simple processes adapted to natural scene properties describe, explain and predict behavioural acts by? This study investigates the hypothesis that human Regions of Interest (ROI) in natural scenes may be predicted from the image structure and from early visual processing.

Our knowledge of the visual system allows for linking the structure to the function and to the behaviour. The local retina adaptation to luminance results in a chromatic invariance [1] and a local contrast enhancement [2]. The magnocellular pathway supports segmentation from movement and local variance [3]. Next, the lateral geniculate nucleus is involved in information gain control. The primary visual cortex could support various phenomena such as illusory contour, pop-out and contextual modulation in target detection [4]. Within this work, we explore how early visual processing may account for eye movements.

A region of interest is a zone of a natural scene exhibiting properties different from those of its neighbourhood. Interest or saliency will be defined for each image through ocular fixations (Human Interest) and through a simple model of the first visual processing (Artificial Interest or Saliency).

1.1 Human Interest

In a behavioural view, selecting ROI means selecting visual information using attentional processing [5]. Attention is a complex behavioural function modulated by stimulus properties, subjects' intentions and experimental context. However most researches about eye movements measure the effect of semantics on eye movement through a manipulation of the experimental instruction [6] or of semantic congruency between objects and context [7]. In this study, we focus only on the signal (or physical) part of the picture (e.g. the pixel luminance). We defined selective attention through an eye movement’s analysis because. Eye movements belong to a sub-class of external manifestations of visual attention. This is why we operationalize selective attention trough eye movement’s analysis. Then, we obtain a simple measure of interest (position and fixation duration) which we compare to different implemented models and statistical tools.

1.2 Artificial Interest

Artificial interest is an approximation of visual processes based on physiological principles. The superior colliculus, a sub-cortical nucleus, triggers ocular saccades [8]. This nucleus receives its afferent connexions from the main oculomotor area and from the striate cortex [9]. Therefore, the first visual process (from retina to primary visual cortex) may trigger a saccade. The early visual processes are identified, defined and simulated by retinal [2, 3] and cortical cells [10] for 10 years. We have built a saliency map [11-13] with *pseudo-physiological* tools (retina and cortical filters) as features or primitives. A saliency map is a representation of the visual field, coding interest or attractiveness on each pixel. In the original model of Koch and Ullman [11] saliency is built from a multi-scale analysis of colour, luminance and orientation. Milanese [12] added symmetry and curvature. Moreover, different works converge on a topographic map for pop-out phenomenon [14], image recognition [15] and attentional selection [16].

2 Model: Saliency Map

The minimal model, we implement (fig. 1), is feed-forward, including the retina and the primary visual area (V1 or Area 17). Neither colour nor temporal information is used in our experiment since we only focus on the spatial structure of natural scenes.

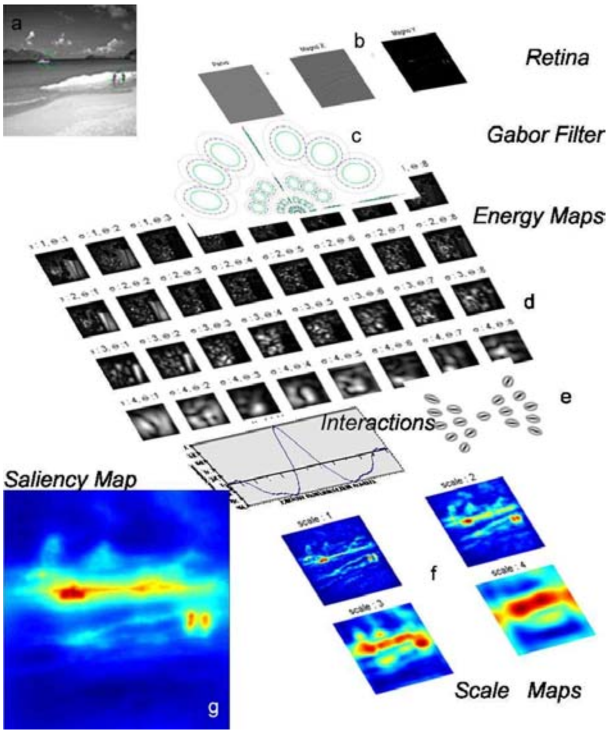

Figure 1: Saliency map model. a. Image, b. Retinal output, c. Bank of filters, d. Energy map, e. Interactions, f. Scale maps, g. Saliency map.

Each picture is processed by a model of the retina, [2, 3] then filtered by interacting Gabor filters, merged in scale maps and frequency maps and then in a saliency map.

2.1 Retina and cortex

When processing an image, the cones and their synaptic triads adapt to luminance which results in a chromatic invariance [1] and a local contrast enhancement. [2] Thereafter the parvocellular ganglion (fig. 1b) cells provide a high-pass filter known to whiten image's frequency spectrum (that is compensating the natural images of 1/f spectrum) [3].

The outputs of ganglion cells are filtered by a set of Gabor filter (simulating cortical simple cells). The outputs of filters are used to build the feature or energy maps (fig. 1c) by coding the local energy spectra (simulation of the complex cells - equation 1 & fig. 1d). Gabor functions are used because they "fit the 2D spatial and spectral structure of simple cells in visual primary cortex, with a small non-structured error indiscernible from random error" [17]. The parameters of the filter banks are chosen to fit the psychophysical data of anisotropy [18] in visual search between vertical and 18 deg-oriented bars.

$$E_{f_c,\theta} = I(x,y) * G(x,y,\sigma_x,\sigma_y,f_c,\theta) \qquad (1)$$

A Gabor filter is made of a Gaussian (with parameters σ_x and σ_y for its spatial extent) modulated by a complex exponential with frequency f_c and direction θ.
We have five frequency bands f_c=[0.3, 0.1306, 0.0568, 0.0247, 0.0108], with the corresponding spatial extents σ_x = [2, 4, 9, 22, 52] and $\sigma_{y=}$[2, 5, 12, 28, 65] in pixels, and eight orientations, θ =[0, 22.5, 45, 67.5, 90, 112.5, 135, 157.5], in degrees.

2.2 Interactions

A set of interactions is defined between and within the feature maps (fig. 1e).

The first ones are inspired from works on target detection [4] and on structure of dendritic and axonal arborisation [19] which lead to "association or extension fields" [20]. An association field is a region surrounding the receptive field of a neuron: "the association field will influence the discharge evoked by a test stimulation shown in the centre of the receptive field" [19]. They are implemented with a Gaussian windowing a butterfly mask. Each field is adapted to the orientation and the central frequency of each energy map. As a result, it increases contour by emphasizing on collinear and curvilinear filters and decreasing orthogonal ones [21].

The *between* maps interactions are implemented through a linear combination of maps tuned to the same frequency but to different orientations. The weight coefficients are computed [22] in order to simulate the sharpening of cortical orientation columns [23]. The result of these interactions lowers the noise and sharpens the tuning of the maps.

2.3 Saliency Map

The feature maps merge into scale maps (fig. 1f) and are weighted (equation 2) according to the following constraints: cortical cells tuned to horizontal and vertical orientations are almost as numerous as cells tuned to other orientations [24]. This is probably due to the fact that the average spectra of natural scenes have strong horizontal and vertical components (see also [25])

$$M_{f_c}(x,y)=\frac{1}{2}\left(E_{f_c,0}(x,y)+E_{f_c,\pi/2}(x,y)\right)+\frac{1}{\Theta-2}\sum_{\theta\neq 0,\pi/2}E_{f_c,\theta}(x,y) \qquad (2)$$

The scale maps are finally integrated to obtain the Saliency map (equation 3 – fig. 1g, 3b 3d), which may be used to predict the regions of interest [12] or fixation points [13].

$$S(x,y)=\sum_{f_c=1}^{F}M_{f_c}(x,y) \qquad (3)$$

The saliency maps select regions with high energy persisting through the scales and regions that are different from their neighbourhood. The size of the neighbourhood depends on the size of the Gabor filter and of the Gaussian "extension field".

3 Human Interest Map

Once the artificial interest is built, we look for a measure of human interest. The ocular fixations sample natural scenes under signal, semantic and intentional constraints. Therefore, the measurement of eye-movement provides us with a simple measure of interest (position and fixation duration) for each picture.

3.1 Method

3.1.1 Subjects

50 volunteer students aged 18 and 25subjects participated; none had previous experience with oculomotor experiments. All subjects had normal or corrected to normal vision.

3.1.2 Display

The visual stimuli were presented on a computer screen (45 x 30 cm) of a viewing distance of 60 cm. A graphic resolution of 1280 x 1024 pixels and a frame rate of 100 Hz were used. Ninety-six natural images, belonging to 4 categories (beach, city, field and room), were presented to the subjects. All pictures' size was 512 x 512 pixels and they were coded in 256 grey levels. Images' contents were chosen to represent a field of view similar to that of the human one.

3.1.3 Eye movement measurement

The movements of both eyes are measured with a head-mounted infrared reflection device (Eyelink, SMI) with a resolution of 0.01 deg and a theoretical accuracy of 0.5 deg. raw eye positions are corrected by linear interpolation using a 3 x 3 point grid. A point is randomly presented before each block. Within a block, after every tenth image a brief central fixation cue allows a drift correction and an error measurement. The estimated tracking error was at worst, less than 0.9 degree of visual angle. The position signal was digitised with a sampling rate of 250 Hz.
In the on-line analysis, fixations and saccades were detected by a mean of a motion (d > 0.15 deg), velocity (v > 30 deg/sec) and acceleration (a > 9500 deg/sec^2) criterion.

3.1.4 Procedure

The subjects are placed in a darkened room. They are instructed not to move their head and body and not to close their eyes during the experience.
Images are presented in two blocks of 48 images each.
Subjects are instructed to fixate on a blue square (32 x 32 pixels) presented in one of the screen corners in order to trigger the image. A natural scene is displayed on the left or the right of the screen (counterbalanced between subjects) during 3 sec. The blue target and the picture do not overlap in order to avoid any bias due to a first fixation on the image.
Subjects announce the category of each image in a microphone. We use a verbal answer to ensure that the subjects stay involved during the entire experiment.

3.2 Data and Results

For each subject recorded, we evaluate an error, according to the time spent before triggering the blue target, the numbers of blinks and the measure of drift.

3.2.1 Data

The mean fixation duration is 300 ms and the mean saccade velocity and angle are in accordance with previous works [7]. There is no difference between categories. Individual scan-paths are ignored because the aim of the experiment is not to study the variability of eye movements but the shared regions of interest.

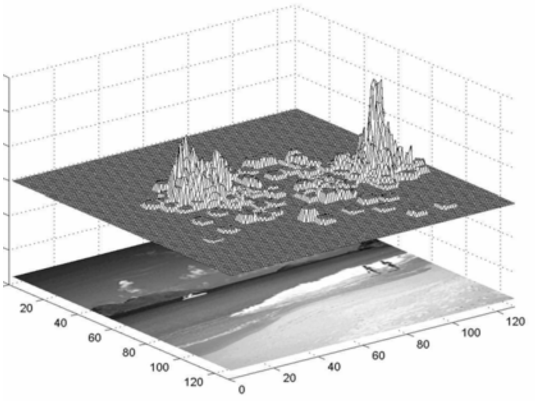

Figure 2: Fixation density map super-and the corresponding scene

4 Comparison

In order to compare ocular recordings and saliency maps, we build a density fixation map (fig. 2) averaging data from all subjects on each image [26, 27]. We increment the density maps within a circular region of 1 dg size (foveal region) for each fixation with the value of the fixation duration. With this, we obtain a 3D map (fig. 2) where the third dimension codes for the time spent observing an area, which is a measure of human interest for each pixel. We use level curve (fig. 3a, 3c). — Contour plot extracted from the 3D saliency map for a particular threshold — to select gaze-attractive regions. A first and rough approximation showed close correlation between saliency maps and fixation density maps as they selected similar areas (fig. 3).

A possible evaluation of similarity between the artificial and the human region of interest is found by computing saliency and other statistical descriptors in foveal regions.

4.1 Images Descriptors

In order to test the hypothesis that the raw signal is an attractor giving rise to eye movement in natural environment, we evaluated the signal in foveal regions with saliency and luminance descriptors. We suppose that the statistical properties of regions selected by eye movements are remarkable or different from the rest of the picture. Therefore, we compute negentropy (the opposite of entropy), contrast, variance and normalised contrast as Reinagel and Zador [28 2911]. We have also measured saliency and scale maps value for those regions.

Entropy (equation 4) measures richness and diversity inside a region. This descriptor has its maximum for a flat histogram (each pixel has the same probability (number of pixels / number of grey level) to have any luminance value) and its minimum for a distribution where all observations fall in one bin (every pixel has the same luminance). Unlike saliency and perhaps variance (Y magnocellular ganglion cells [2]), entropy could not be an ecological measure or plausible from a biological point of view. We also measure normalised contrast: normalised contrast "refer to the local standard deviation within a patch, normalized by the global mean intensity of the image which simply reflects the variance of the local pixel intensities, on the spatial scale of the size of the fovea." [28].

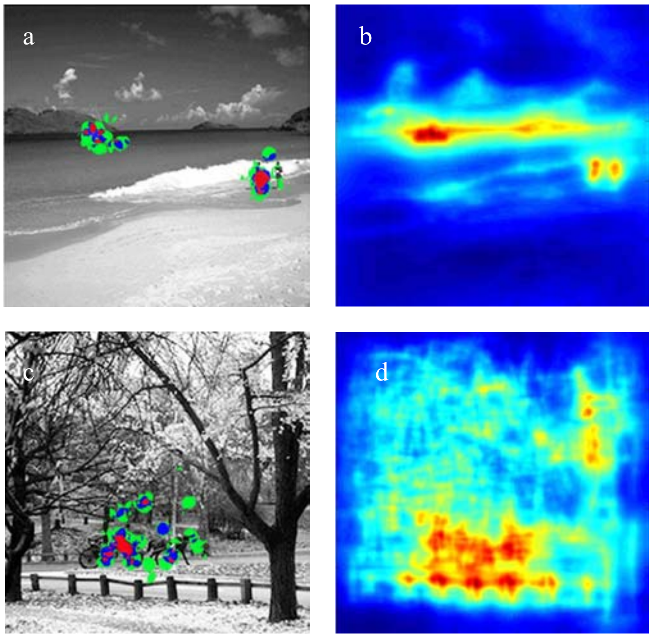

Figure 3: Human and artificial interest. Figures a and c show regions selected with density fixation maps superimposed on the original picture. Figures b and d show saliency maps.

Descriptors from eye-selected regions are compared to a systematic sampling instead of a random one in order to avoid any bias of the size of the sample. The systematic inspection means that the entire image is explored through a sampling grid of patches sustaining one square degree. All descriptors values are averaged for each image giving a systematic value per image which is the boundary value for the random sampling.

In order to control and evaluate the interest of our model we compare every descriptor to the saliency and the scales maps. Each descriptor value is centred and reduces according to the mean and the standard deviation of the systematic samplings.

$$H = \sum_{i=1}^{\max(I)} \rho_i \log(\rho_i) \qquad (4)$$

$$C = \frac{Max(x_i) - \min(x_i)}{Max(x_i) + \min(x_i)} \qquad (5)$$

$$\sigma = \frac{1}{\bar{I}} \sum_{i=1}^{N} (x_i - \bar{I})^2 \qquad (6)$$

$$CN = \frac{1}{N\bar{I}} \sum_{i=1}^{N} (x_i - \bar{I})^2 \qquad (7)$$

With x_i Pixel i intensity, N Number of pixels in each region, ρ_i probability of a pixel to belong to the luminance class i (The histogram is normalised to become a probability density function $\Sigma\rho_i=1$).

$$\bar{I} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (8)$$

4.1.1 1st Results

As shown in figure 4, box plots evaluate the significance of differences between the two modes of region selections (human vs. systematic).
Box plots show the mean and variance for each measure. As Reinagel and Zador [28], differences are significantly higher for each subject when compared to systematic measures, except for the variance.
When comparing the descriptors together, the measures show no difference except for variance and contrast which are significantly lower than the others.

4.1.2 Variable Resolution

When the visual system selects eye's next landing point, the resolution outside the fovea is highly reduced. The acuity is maximal in the central degree and decreases with the eccentricity. This is due, among other, to the morphology of the crystalline lens, the distribution of receptors, and the convergence of the retinal networks causes. If signal is an attractor, the descriptor values will be higher for the current region fixated (fig. 4) and for the following region.
Therefore, we computed statistics for the regions following each fixation according to the simulated resolution of the distance between fixation and next landing point (except for saliency because it is already a multi-scale analysis). The acuity and the resolution are computed using the Sére et al model [29]:

$$A(e) = \frac{1}{1 + 1.87e} \qquad (9)$$

e =angular eccentricity in degree

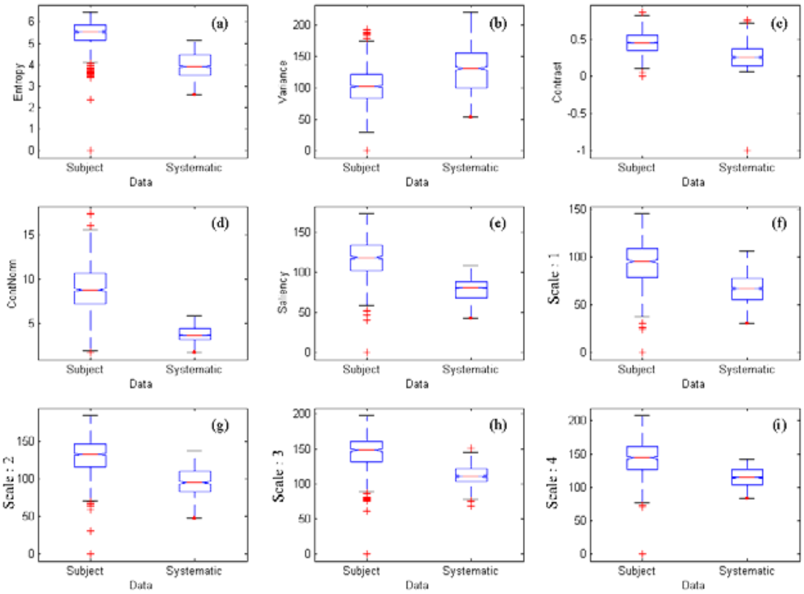

Figure 4: Box plots of a) Entropy, b) variance, c) Contrast, d) Normalized contrast, e) Saliency, and 4 scale maps: f) High frequency to i) Low Frequency

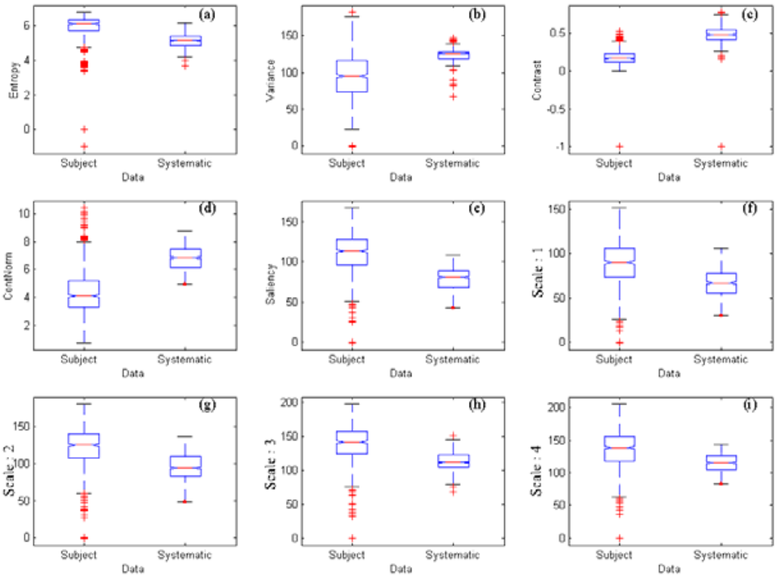

Figure 5: Box plots of a) Entropy, b) variance, c) Contrast, d) Normalized contrast,with variable spatial resolution and e) Saliency, and 4 scale maps: f) High frequency to i) Low Frequency

4.1.3 2nd Results

As shown in figure 5, statistical values are not significantly higher for regions selected by subjects than for the control condition (systematic) except for entropy. For contrast and normalised contrast, these values are inferior to the systematic inspection.

5 Discussion

We have shown that it is possible to simulate a simple behaviour of gaze orientation in an ecological condition by selecting interest regions using a simple model based on the first stages of visual processing (retina, primary cortex),.

We replicate the findings of Reinagel and Zador showing “that active selection affected the statistics of the stimuli encountered by the fovea” [28]. But, when one takes into account the eye non-homogeneous resolution, entropy and saliency are the only descriptors that predict or correlate with the selected regions. However, if one consider that, unlike saliency, entropy is not a plausible descriptor from a biological viewpoint, our results valid the idea that saliency is a relevant descriptor for regions of interest in natural scenes.

Nevertheless, the descriptors used were based on luminance properties only and could not account for any semantic properties, although in natural environment meaning and signal are inseparable. Every picture is made up of a set of objects belonging to the meaningful world and of a set of grey levels pixels. Roughly speaking, each pixel is projected in a physical/signal space and in a semantic space. Therefore, in natural scenes, every region is meaningful. Looking back to obtained scan-paths, we can see that fixations often fall on objects or on objects’ parts. So, the selected regions could be selected because they are significant or, as shown, "salient". Therefore, our results do not allow us to posit that the sole saliency account for the human selected regions of interest. In further experiments, we will manipulate independently the physical and semantic properties of objects in order to evaluate their relative weight in visual attractors.

6 Acknowledgements

This research was funded by Pierre Mendès-France and Joseph Fourier Universities, Région Rhône-Alpes and Centre National de la Recherche Scientifique (UMR 5105).

References

1.Alleysson, D. and J. Hérault. *Differential thresholds in colour perception: a consequence of retinal processing and photoreceptor non-linearities.* in *European Conf. on Visual Perception*. 1998. Oxford UK.

2.Beaudot, W., *Le traitement neuronal de l'information dans la rétine des vertébrés : Un creuset d'idées pour la vision artificielle*, in *Sciences Cognitives*. 1994, Institut National Polytechnique: Grenoble.

3.Hérault, J., W. Beaudot, and A. Oliva. *Perception Coarse-to-fine par un modèle de rétine.* in *GRETSI*. 1995.

4.Polat, U. and D. Sagi, *The architecture of perceptual spatial interactions.* Vision Res, 1994. **34**(1): p. 73-8.

5.Olshausen, B.A. and C. Koch, *Selective Visual Attention*, in *The Hanbook of Brain Theory and Neural Networks*, M. A, Arbib., Editor. 1998, The MIT Press: London. p. 837-840.

6.Yarbus, A.L., *Eye movment and Vision*. 1967, New York: Plenum Press.

7.Henderson, J.M. and A. Hollingworth, *High-level scene perception.* Annu Rev Psychol, 1999. **50**: p. 243-71.

8.Trappenberg, T.P., et al., *A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus.* J Cogn Neurosci, 2001. **13**(2): p. 256-71.

9.Kandel, E.R., J.H. Schwartz, and T.M. Jessell, eds. *Principles of Neural Science*. 4 ed. 2000, McGraw-Hill: New York.

10.Jones, J.P. and L.A. Palmer, *An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex.* J Neurophysiol, 1987. **58**(6): p. 1233-58.

11.Koch, C. and S. Ullman, *Shifts in selective visual attention: towards the underlying neural circuitry.* Human Neurobiology, 1985. **4**: p. 219-227.

12.Milanese, R., *Extraction de régions saillantes dans un image : de l'évidence biologique à l'implantation sur ordinateur*, in *Informatique*. 1993, faculté des sciences de l'université de Genève.: Genève.

13.Itti, L., C. Koch, and E. Niebur, *Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. **20**(11): p. 1254-1259.

14.Treisman, A. and G. Gelade, *A feature-integration theory of attention.* Cogn. Psychol., 1980. **12**: p. 97-136.

15.Deco, G. and B. Schurmann, *A neuro-cognitive visual system for object recognition based on testing of interactive attentional top-down hypotheses.* Perception, 2000. **29**(10): p. 1249-64.

16.Wolfe, J.M., K.R. Cave, and S.L. Franzel, *Guided search: an alternative to the feature integration model for visual search.* J Exp Psychol Hum Percept Perform, 1989. **15**(3): p. 419-33.

17.Jones, J.P. and L.A. Palmer, *The two-dimensional spatial structure of simple receptive fields in cat striate cortex.* J Neurophysiol, 1987. **58**(6): p. 1187-211.

18.Ballaz, C., A. Chauvin, and C. Marendaz, *Anisotropie et recherche visuelle : l'orientation canonique comme déterminant de la saillance perceptive.* In Cognito, 2001.

19.Chavane, F., et al., *The visual cortical association field: A Gestalt concept or a psychophysiological entity ?* Journal of Physiology, 2000. **94**: p. 333-342.

20.Petitot, J. and Y. Tondut, *Géométrie de contact et champ d'association dans le cortex visuel*. 1998, École Polytechnique: Paris.

21.Kovacs, I., *Gestalten of today: early processing of visual contours and surfaces.* Behav Brain Res, 1996. **82**(1): p. 1-11.

22.Carandini, M. and D.L. Ringach, *Predictions of a recurrent model of orientation selectivity.* Vis. Res., 1997. **37**: p. 3061-3071.

23.Somers, D.C., et al., *A local circuit approach to understanding integration of long-range inputs in primary visual cortex.* Cereb Cortex, 1998. **8**(3): p. 204-17.

24.De Valois, R.L. and K. De Valois, *Spatial Vision*. 1988, Oxford: Oxford University Press.

25.Field, D.J., *Relations between the statistics of natural images and the response properties of cortical cells.* J Opt Soc Am [A], 1987. **4**(12): p. 2379-94.

26.Mannan, S.K., K.H. Ruddock, and D.S. Wooding, *Fixation sequences made during visual examination of briefly presented 2D images.* Spat Vis, 1997. **11**(2): p. 157-78.

27.Mannan, S.K., K.H. Ruddock, and D.S. Wooding, *The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images.* Spat Vis, 1996. **10**(3): p. 165-88.

28.Reinagel, P. and A.M. Zador, *Natural scene statistics at the centre of gaze.* Network, 1999. **10**(4): p. 341-50.

29.Séré, B., C. Marendaz, and J. Hérault, *Nonhomogeneous resolution of images of natural scenes.* Perception, 2000. **29**: p. 1403-1412.