Título: MODELING EARLY VISUAL CODING AND SALIENCY THROUGH ADAPTIVE WHITENING: PLAUSIBILITY, ASSESSMENT AND APPLICATIONS

Nombre: García Diaz, Antón

Universidad: Universidad de Santiago de Compostela

Departamento: Electrónica y computación

Fecha de lectura: 15/04/2011

Programa de doctorado: Progr. Interuniversitario en Tecnoloxias da Información

Dirección:

> **Codirector:** Xosé Manuel Pardo López

> **Codirector:** XOSE R. FERNÁNDEZ VIDAL

Tribunal:

> **presidente:** Diego Cabello Ferrer

> **secretario:** Filiberto Pla Bañón

> **vocal:** JAVIER MARTÍNEZ BAENA

> **vocal:** Xoana Gonzalez Troncoso

> **vocal:** JOSE M CAÑAS PLAZA

Descriptores:

> VISION ARTIFICIAL

> TRATAMIENTO DIGITAL DE IMAGENES

El fichero de tesis ya ha sido incorporado al sistema

> 2011garcimodel.pdf

Localización: BIBLIOTECA XERAL USC

Resumen: Biological vision establishes a wide variety of unrivaled benchmarks in terms of efficiency, robustness, and general performance in active visual tasks. Despite the complexity and variability of natural images, visual systems of mammals are surprisingly skilful in recognizing objects and contexts at a first glance and to efficiently drive few fixations to the most salient parts of a new unknown scene.

These capabilities demand an active and dramatic selection of information that poses a main cause for visual attention. It seams reasonable considering the huge flow of information entering the human visual system (HVS) through the retinian photoreceptors, estimated to be over 10^{10} bits/s (Anderson 2005). Bottom-up adaptive processing and perception of saliency, are thought to lie at the basis of this early behavior with such a remarkable efficiency. They appear to play an essential role in the control of human visual attention -working in cooperation with top-down control- as a number of results from a wide variety of experiments have shown.

Visual saliency is usually employed to refer measures that aim to quantify the conspicuity or distinctiveness of a visual stimulus. That is, it intends to quantify how much a stimulus stands out from the context, given its physical properties. The common representation of saliency is given in the form of a retinotopic map (the saliency map). A main -though in no way the only- source of information to understand the functioning of visual attention is the spatial distribution of human eye fixations obtained in eye-tracking experiments. Eye movements result in fixations that determine the small regions of a given image that are sensed by the fovea. Under good illumination conditions (i.e. for photopic vision), these small regions receive a much higher spatial resolution due to the much higher density of photoreceptors present in the fovea. Consequently, eye movements represent a first form of strong spatial selection of visual information. It must be noticed however that also peripheral vision is affected by attentional selection, without the need of eye-movements. This issue will be considered further along this dissertation.

Otherwise, mechanistic models of early visual processing with a biological concern are focused on the explanation of the visual receptive fields and their adaptive behavior, to local and contextual features. A main goal of these models is the formulation of early coding strategies that are biologically plausible and that are able to explain observed visual phenomena related to early vision, and particularly to contextual adaptation of perceptual and neural behavior.

The problem of measuring the saliency or distinctiveness in an image has also a great relevance in computer and machine vision, specially in the development of active systems. Indeed, bottom-up spatial attention has shown to be very useful in important visual functions like learning and recognition and many vision applications as shown in the first chapter of this thesis. Besides, the extraction of suitable low level features is of enormous importance in image analysis and computer vision. Both of these problems -low level representation and saliency- use to appear closely related in a variety of solutions. A remarkable example can be found in the most popular interest point detectors, but also in many other computer vision models.

Both concerns on the understanding of the HVS and on the development of active vision systems have fostered an important and crossdisciplinary research effort to provide improved measures of saliency. Particularly, the bioinspired modelling of saliency and its applications have seen an extraordinary and increasing amount of research efforts in the last years.

However, there is clearly a lack of models that address the relationship between the contextual data-driven adaptation observed in early visual coding and the perception of saliency. Understanding this relation is essential for the development of a computational framework of early visual coding with biological plausibility. Such a framework should formulate plausible intermediate retinotopic representations adapted to the image. These intermediate representations must be able to maintain a suitable measure of saliency, but also to match observed characteristics of early vision. Approaches to this problem are very interesting for computer vision too, as far as they may yield improved models of both adaptive low level features and saliency.

Furthermore, most models of saliency are grounded on an information theoretic foundation, without an specification of the physical sources involved, and more importantly, of the different ways in which they contribute to visual saliency. This specification, if possible, is very important since it would offer an additional constraint to understand the visual function in terms of its physical roots. As well it could yield excellent cues for

the development of active vision approaches and in general for the adaptive processing and analysis of images.

With the aim of filling these gaps, this thesis provides a coherent functional approach to both early visual coding and saliency, in a biologically plausible manner. Likewise, the framework proposed is rooted in a physical interpretation involving few simple optical magnitudes. The resulting model is shown to explain a variety of visual illusions and to clearly outperform the existing state-of-the-art models of saliency using the most popular evaluation tests, including the prediction of eye fixations and the reproduction of psychophysical results.

The first pointed lack can be easily appreciated in the two typical strategies of low level representation adopted by existing models of saliency. Many of them start with multiresolution decomposition of three predefined color components, in a given color model. This is done by projecting the image color components on linear filters resembling receptive fields of cells in V1, which are usually modeled by Gabor-like and Gaussian-like functions ever since the standard model of V1 was first proposed by Hubel and Wiesel (1959, 1968). The following steps generally involve a competition and integration process that delivers a final measure of saliency, a scheme already found in early models based on the Koch and Ullman architecture of attention (1985). Otherwise, the other typical approach involves decomposition through the projection of the image on independent components of natural image patches, avoiding color components and filter parameterization beyond patch size. This proposal is based on the statistical interpretation of the standard model of V1 as the result of evolution and neural development to match the statistics of natural images (Olshausen 1996, Bell 1997).

Both of these schemes, either based on filter banks or on independent components analysis, share an important property: they always use the same portions of the feature space to represent any image. Filter bank approaches project a fixed set of color components on a fixed partition of the spectral domain. Independent components are determined from a set of training natural images and are not modified subsequently.

The described static approaches to early coding underlying most of current models of saliency do not match the behavior of the HVS. Indeed, it adapts its responses to the global and local features of each specific image. It shows short-term and contextual adaptation to contrast, to color content and to spatial structure. This adaptation takes place from photoreceptors and G cells to cortical cells and has been shown to produce overall a decorrelated representation (Barlow 1989, Rieke 2009, Kohn 2007, Clifford 2007, Schwartz 2007). Adaptive decorrelation seems thus to be a plausible neural mechanism. Not surprinsingly, many recent mechanistic models of neural cortical networks as well as models of computation by populations of neurons produce an overall decorrelated and whitened representation of the input.

From a computational point of view, there are also reasons in favor of a contextual adaptation model. Approaches that do not present such adaptation are more likely to be affected by feature biases, reducing the applicability of the corresponding measure of saliency.

Therefore, the problem of saliency appears to be closely related to the problem of selection of a low level representation as well as its adaptation. In the context of biological vision, early visual coding appears to be an unavoidable problem to tackle wether biological plausibility is claimed. Otherwise, a proper insight in early visual coding can deliver guidelines to design low level representations of images, suitable for active visual functions that migth be useful for computer and machine vision applications. Indeed, and similarly to other works in the

field, the original motivation of this dissertation was born within a long-term project of developing a generic and biologically inspired framework to approach and study active vision problems.