Universitat
Autònoma
de Barcelona

Master in Computer Vision and Artificial Intelligence
Report of the Master Project
Option: Computer Vision

Exploring low-level vision models. Case study: saliency prediction

Author: **Ivet Rafegas Fonoll**
Advisor: **Dr. Maria Vanrell Martorell**

Acknowledgements

I would like to thank my advisor Dr. Maria Vanrell for her ideas, motivation and proposals. She has been providing me all the information that I needed and has been giving me the required explanations, as well as good proposals for structure this dissertation appropiately. I would also like to thank Dr. Xavier Otazu for the induction model bases explanation, as so as for providing me the necessary code implementation.

Special thanks to all my family for their support. To Roger, for his support, patience and for trusting me in every moment.

Thanks to Quim for his patience and helping me in everything I needed while living together. I wish I could share more moments with him.

I would like to thank all my colleagues at the Computer Vision Centre, C. Sánchez, A. Hernández, L. de las Heras, J. Núñez, J. Almazán, D. Fernández, C. Davesa, F. Cruz, Y. Socarrás, R. Balagué, A. Dutta,... for their encouragements towards me when I did not know how to continue and also for the nice moments we shared in the centre.

Special thanks to Gemma, her outstanding english has been really helpful in my master thesis.

Thanks to Mònica, Mercè and Marina, for cheering me up and for their worrying about my thesis development.

This work has been supported in part by project TIN2010-21771-C02-01.

ABSTRACT

After a decade of huge progress in computer vision using flat processing schemes, new architectures based on deep hierarchies [1] are currently gaining strength in the field. This new paradigm is in line with how visual information is hierarchically processed in the human visual system [2]. Several low-level vision models have proposed hierarchical schemes to simulate the first stages of the ventral stream. In this work, we analyse three of these models (ID, HMAX, MP) giving a unifying overview of them. In this project we focus in what we call the Induction-Derived (ID) model ([3], [4], [5], [6]). The choice is based on its generalisation properties shown in predicting both colour induction effects and saliency maps. Its main stages can be summarized as a linear filtering (L1) followed by a centre-surround mechanism (L2), a divisive normalisation (L3) and the application of a weighting function (ECSF) (L4). Our aim is exploring each layer function by proposing alternative implementations to achieve more accurate responses. As case study, we work on saliency prediction since a large standard datasets are available allowing testing the effects of the studied alternatives. We analyse the DOOG filter family vs. the multi-resolution wavelet, a shaped centre-surround vs. the previous rectangular window different divisive normalisation functions vs. the rational quadratic, and null or random weighting function vs. the ECSF. We conclude that extending the family of filters improves feature selectivity; shaped centre-surround can provide a more accurate response and that divisive normalisation functions require being better fitted to the task in hand. After analysing performance measurements of saliency prediction we propose a new measurement, WARP, and we compare and evaluate all the proposed alternatives in a large set of experiments that provides with a wide analysis of the L1, L2 and L3 model stages. Performed experiments appear to diminish the effects of L4 weighting stage (ECSF). Additionally, we start to explore how we can scale on these hierarchies, in view of a more complex task such as object recognition. We derive a new representation from the model output that can be the starting point for a trainable layer that could give a visual code for object recognition.

Keywords: *bio-inspired, centre-surround, deep hierarchies, divisive normalisation, early vision models, HMAX, induction, Malik-Perona, saliency estimation, saliency estimation measurement, saliency map, ventral stream, visual codes, visual cortex, V1-like*

Contents

Chapter 1

Introduction

In this chapter we motivate our work by adopting a computer vision methodology that takes inspiration from the processing levels of the human visual system. We propose to explore previous low-level vision models in order to learn how they work and hypothesise about specific improvements, having in mind its extension to higher-visual tasks as object recognition.

1.1 Motivation

Primate visual systems shows what seems an effortlessly ability to recognize objects. Understanding its procedure became a target in several research fields in the last decades, ranging from psychophysics, neurophysiology, neuroanatomy, cognitive and computational neuroscience to computer vision [7]. After a decade of huge progress in computer vision using flat processing schemes, new architectures based on deep hierarchies [1] are currently gaining strength in the field, since a lot of evidences conclude that it is performed by the ventral stream pathway that presents a hierarchical architecture. This new paradigm is in line with how visual information is hierarchically processed in the human visual system [2]. Although this hierarchy way is recent, there are several works providing support to this idea ([8], [9], [10], [11]). Figure 1.1 shows schematically the difference between the two frameworks proposed in computer vision. In flat processing, the algorithms were built specifically to solve a task from some kind of features. In deep hierarchies, each layer codifies its inputs to a new representation which is interpreted by the next layer successively until the the last one, which transfers its responses to the task-adapted part. Therefore, separate information channels are successively combined building paths that are progressively more complex and invariant. This fact favour the computational efficiency, due to features are simplified in different channels depending on the characteristic (color, shape, ...) [2]. It seems a good path to follow in future researches to get computational algorithms closer to biological procedure. In this new line of modelling, several layers interact to incorporate successively responses selectivity and tolerance to the identity-preserving. The adaptation to the specific visual task is, then,

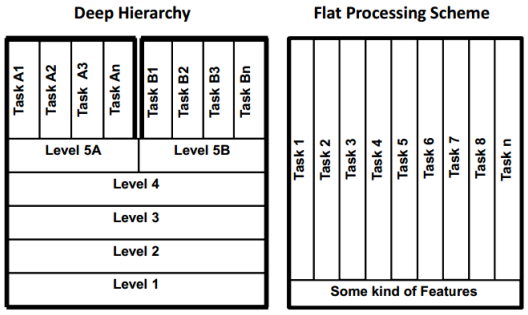


Figure 1.1: (Image from [2]). Layer comparison between deep hierarchies and flat processing.

a post-processing.

Image perception is carried through the visual cortex (from retina to IT, through V1, V2 and V4) (see Figure 1.2). Several low-level vision models have proposed hierarchical schemes to follow the first stages of the ventral stream (where all representations keep the location information encoded), in particular, in V1. This might be due to the ignorance of the brain running in these last steps. However, the hierarchical architectures provided by bio-inspired low-level vision models allow to study the abstractions provided by the intermediate levels. These abstractions get an interesting methodology towards the definition of visual codes for object recognition, which follow somehow what occurs in the Inferior Temporal cortex (IT) [7]. The term visual codes refers to a set of responses representation that our human visual system is able to process and recognise in last stages of our cortical processing, in spite of their complexity. Nevertheless, the authors of [12] also emphasise the needed of building computational algorithms in the visual codes line where usual machine learning techniques can be used to achieve the understanding of the perception.

Following with the algorithms that model early stages of the visual system, the authors of [7] refer them as normalized linear-non-linear (NLN) due to the alternation of linear, non-linear, and normalization operations that they apply to the image perceived (retina's observation). This overview points out that most bio-inspired models follow the same framework. In fact, recent researches are focused in the *V1-like* representations [13], which are computationally built using a spatial linear filtering as Gabor wavelet and combining these responses with some threshold, saturation and normalisation [13]. Therefore, V1-like features are obtained using a NLN algorithm. In our work we will study three low-level vision models: Induction-derived family of model (ID) ([4], [3], [5], [6]), HMAX algorithm [8] and Malik-Perona's model (MP) [14], relating them to the NLN categorization.

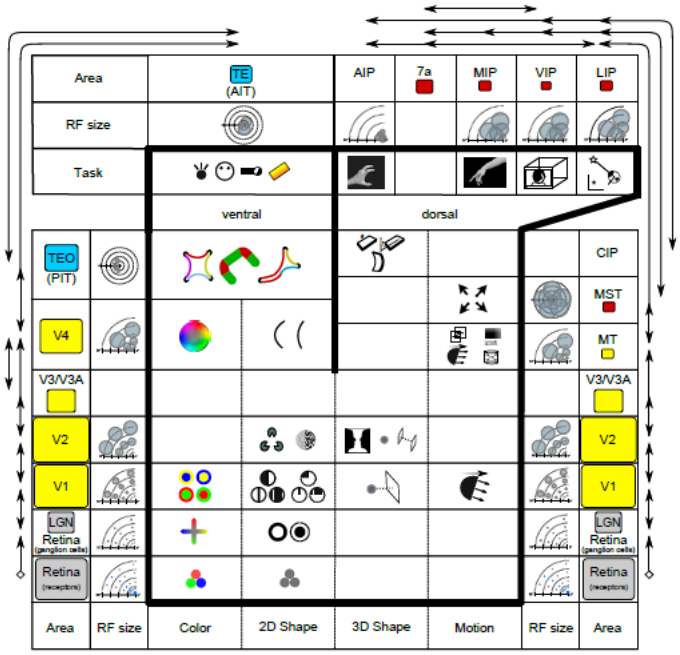


Figure 1.2: (Image from [2]). Overview of the visual processing over different areas of the human visual system.

1.2 Objectives

The objectives of this dissertation can be summarised as follows:

- To review previous low-level vision works analysing commonalities and differences between them. We will select a subset of models, such as ID, HMAX and MP.
- To study the ID model ([3],[4], [5],[6]) in more detail by analysing it stages and exploring different alternatives to improve the stages. We want to study the effects of changing the family of linear filters used; the effects of the window-shapes used in the centre-surround measurements; and to study the function used to perform the divisive normalisation, as well as the true importance of the weighting function. In particular:
  - We will analyse the family of DOOG filters vs. multi-resolution wavelet.
  - Compare the effects of using adapted centre-surrounds (shaped) vs. the constraint rectangular (unshaped) derived from the multi-resolution decomposition.
  - Study the effects of using a parametric sigmoid vs. quadratic rational for the divisive normalisation.
  - Study the effect of removing the weighting function.

- To perform the experiments of this study we are going to work with dataset provided by the problem of estimating visual saliency, since it is giving a good framework to test the models.
- To investigate about how to get closer to the visual codes representation that can be derived from the ID model and try to infer how they can be extended and trained for higher-level visual tasks. We will use SVM to learn a weighting function.

1.3 Report organisation

This work follows with four chapters. First, in chapter 2 we analyse three low-level computational methods which are studied under a layer-division or hierarchy that we propose. Chapter 3 is focused on the analysis of the ID, under the saliency estimation case. It presents an early vision to saliency prediction and explains how we modify it to a more accurate model. Next, chapter 4 discusses about the measurements that are usually used when evaluating saliency estimation and proposes a new metric. Also, this chapter shows a set of experiments run to go in depth in the model, including their results. Finally, chapter 5 proposes new directions to continue and improve our method and expose our conclusions.

Chapter 2

Related work

This chapter is structured in four sections. First, in section 2.1 we propose new V1 layer division to gather three different bio-inspired approaches. Next sections are focused on the layer specification particularly for each algorithm, under our new V1 hierarchy. ID is explained in section 2.2, followed by HMAX in section 2.3 and concluding this chapter with section 2.4 where the Malik-Perona algorithm is described.

2.1 Overview

This section reviews bio-inspired approaches for low-level visual representation. We focus our research on three methods that have been applied to different visual tasks. In one hand, we study the main stages of Induction-Derived family of model (ID) ([4], [3], [5], [6]) since its steps give rise to different methods. Next, we review the HMAX algorithm [8] based on hierarchical structure with MAX-like operations, which is one of the few models that achieves to work on recognition tasks in a feed-forward manner. Although Zhang *et al.* extended HMAX algorithm to incorporate color on it [15], in this work we only study the 2-D shape stimuli (see figure 1.2). Finally, we recover Malik-Perona's model (MP) [14], proposed at the end of the 80's, that is a way well justified model and well-known in the computer vision field. All of these approaches have provided with feed-forward hierarchical architectures whose levels are somehow justified by known neuronal mechanisms of the visual system. In fact, we will see that all these methods make similar operations to the V1-like model.
Starting a review of related works with a conclusion is not common, we prefer to start with a summary of our overview for the models, similar to the NLN categorisation. We believe that it helps to the comprehension of the relation between all the methods. Therefore, before looking thoroughly at each method we propose a global overview to compare all methods. In this dissertation we carry out an abstraction exercise to analyse common and unique steps for all these methods. Thanks to this procedure we notice that they share several mechanisms with the same or similar goals, and we redefined their steps into

five layers that cover all the methods. Table 2.1 shows our bio-inspired layer hierarchical proposal. In the following lines we explain the meaning of each layer, without going into the details of any of the methods.

L1 - Linear filtering. The first layer consists of a frequency-orientation selectivity. It is shared by three methods although they use different ways to filter the image.

L2 - First non-linearity. Is where previous responses are refined. HMAX focuses on refining scale responses, while MP focuses on their sign. In the case of ID, it makes a centre-surround enhancement.

L3 - Second non-linearity. It makes a second non-linearity to extend maximum values and threshold the minimum. To this end, HMAX uses a local dilation and a sub-sampling, ID uses a divisive normalisation while MP a post-inhibition response. In this layer the MP model is finished.

L4 - Weighting. A weighting function in ID is applied to the previous selected responses in order to enhance or discard them depending on the visual problem. In this layer is where image representation is finished using ID.

L5 - Coding responses. HMAX follows with a change on its base representation in this layer, where image is described according to certain visual codes or vocabularies. We have to add here that up this point the information follows a retinal composition, this means that they keep spatial information encoded in their outputs. In this line, we interpret that HMAX takes a step forward due to its new way to encode information which takes place on the last layer.

Layer	HMAX	ID	MP
L1 - Linear filtering	Gabor	Multi-resolution wavelet	DOOG
L2 - First non-linearity	Scale refinement: local maximum over two consecutive bands	Centre-surround enhancement using a local mean	Sign selectivity by half rectification
L3 - Second non-linearity	Local dilation + sub-sampling	Divisive normalisation	threshold + post-inhibition response
L4 - Weighting	-	ECSF	-
L5 - Coding responses	Global maximum	-	-

Table 2.1: Overview of HMAX, ID and MP in several layers. Each row shows common proposals between methods, specifying the operations used to achieve them.

2.2 Induction-Derived family models

By Induction-Derived family (ID) we refer to the common pipeline followed by Brightness Induction Wavelet Model (BIWaM) [3], Chromatic Induction Wavelet Model (CiWAM) [4], Saliency by Induction Mechanisms (SIM) [5] and the extension of SIM using grouplets (SIM+GT) [6]. The name of Induction-Derived comes from the fact that the model was initially built to predict colour induction effects.
ID models are a bottom-up approach, which use a scale-space decomposition thanks to the multi-resolution wavelet transform, the idea of centre-surround differences and a weighting function which became a particular key of the model. In the following paragraphs, we detail all of the steps of ID hierarchy, using the mathematical nomenclature described by N. Murray *et al.* in [5].

L1. Given a channel-image in opponent color space (I_c), it is convolved with a set of Gabor-like basis filters. It achieves a spatial decomposition thanks to the use of multi-resolution wavelet transform[1] (WT) for different wavelet planes orientations (horizontal (h), vertical (v) and diagonal (d)).

$$WT(I_c) = \{w_{s,o}\}_{s=1,2,...,n;o=h,v,d} \tag{2.1}$$

where I_C is one of the image opponent channel representations, $w_{s,o}$ is the wavelet plane at a certain scale s and orientation o, assuming n different scales.

L2. This step is where the concept of centre-surround gains relevance and tries to model the effect of the casing to a certain region. For each position in each wavelet transformation ($w_{s,o}$), they calculate a local mean in a certain neighbourhood according to its orientation o and scale s. Using the mean measurement, they are able to obtain the contrast centre energy ($a^{cen}_{s,o}(x,y)$) and the contrast surround energy ($a^{sur}_{s,o}(x,y)$) for each $w_{s,o}(x,y)$ wavelet coefficient centred at position x, y. To estimate the interaction between the centre and surround regions, the authors of [3] suggest the use of the following equation:

$$r_{s,o}(x,y) = (a^{cen}_{s,o}(x,y))^2/(a^{sur}_{s,o}(x,y))^2 \tag{2.2}$$

L3. Here is where the concept of divisive normalisation appears. D. Heeger reports that "complex cells have several linear parts that are rectified before being combined into the complex cell response" [16]. One of the main problems of linear or energy models is the fact that cell responses saturate at high contrasts. Therefore, the idea of applying a divisive normalisation is to enhance some contrast from the others. ID has a centre-surround energy measurement ($z_{x,y}$) that follows the idea of the divisive normalisation concept using a quadratic rational function:

$$z_{s,o}(x,y) = r^2_{s,o}(x,y)/(1 + r^2_{s,o}(x,y)) \tag{2.3}$$

[1] The pyramid-scale decomposition is achieved thanks to the power of two reshape diminution (down-sampling).

L4. ID has a weighting function based on the Contrast Sensitivity Function (CSF) of Mullen [17] to rectify energy representations according to human perception. The idea of applying this weighting function is to discard or emphasise some energy responses. X. Otazu *et al.*, in [4], extended CSF (ECSF) and it is described in terms of scale (s) and centre-surround (z) parameters. ECSF was determined adjusting two gaussian functions by psychological experiments and was subsequently readjusted for saliency by least square regression to determine parameters of ECSF [5].
Therefore, each energy measurement ($z_{x,y}$) is modified by ECSF function as follows:

$$ECSF(z,s) = z \cdot g(s) + k(s) \tag{2.4}$$

where

$$g(x) = \begin{cases} \beta \exp -\frac{(s-s_0^g)^2}{2\sigma_1^2} & s \leq s_o^g \\ \beta \exp -\frac{(s-s_0^g)^2}{2\sigma_2^2} & otherwise \end{cases} \tag{2.5}$$

$$k(x) = \begin{cases} \exp -\frac{(s-s_0^k)^2}{2\sigma_1^2} & s \leq s_o^k \\ 1 & otherwise \end{cases} \tag{2.6}$$

where σ_1 , σ_2, σ_3 , β, s_0^g , and s_0^k are adjusted parameters.

To sum up, in this step they use $ECSF$ function to weight the centre-surround contrast energy $z_{s,o}(x,y)$, and will become the image representation in this model:

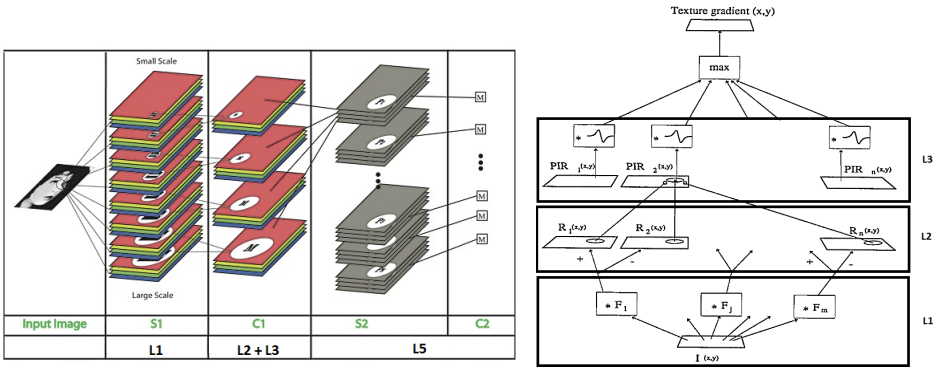
$$\alpha_{z_{s,o}}(x,y) = ECSF(z_{s,o}(x,y), s) \tag{2.7}$$

2.3 HMAX Model

T. Serre *et al.* [8] defined a general bio-inspired method for object recognition in visual tasks. The model is based on a theoretical model of the feed-forward processing of object recognition in ventral stream [18], which holds invariance and selectivity to be the main properties of the recognition task. The framework defined by T. Serre *et al.* is known as HMAX, since it is based on a hierarchy of MAX-like operations ([19],[20]) as might be found in the cortex.

HMAX framework alternates a template matching and a maximum pooling operation which allow a good trade-off between selectivity (*S units*) and invariance (*C units*), respectively. These stages are graphically shown in figure 2.1(a), but we will separate each step according to our layer nomenclature. Below are described the main stages of HMAX, using a new mathematical formulation which is not provided in [8].

L1. This layer corresponds to the simple cells found in the primary visual cortex (V1) that T. Serre *et al.* labelled as *S1 units*. A gray-scale image (I_{gray}) is filtered by a set of Gabor filters with different



(a) (b)

Figure 2.1: (Images from [8] and [14]). Schemes of HMAX (a) and MP (b) related with our layer division

scales and orientations:

$$I_{gray} \longrightarrow S1_{s,o} = G_{s,o} * I_{gray} \tag{2.8}$$

where $G_{s,o}$ is the image response for each Gabor filter in a certain scale s, and orientation o.

L2. This layer performs a local maximum operation between image responses that belong to the same band and share orientation and correspond to the first step of the *C1 units* defined in [8]. T. Serre *et al.* define a band as a couple of adjacent $S1_{s,o}$: band $b_{s'}$ contains $\{G_{s_j,o}\}$ and $\{G_{s_{j+1},o}\}$ $\forall o$, where $j = 2s' - 1$. Thanks to this maximum operation they are able to reduce the information by half:

$$\left.\begin{matrix} G_{s_j,o} \\ G_{s_{j+1},o} \end{matrix}\right\} \longrightarrow C1_{a_{s',o}} = max(G_{s_j,o}, G_{s_{j+1},o}) \tag{2.9}$$

L3. In order to follow with the construction of the *C1 units*, HMAX takes into account the need for a non-linear operation to apply to the complex cell response. For this purpose, T. Serre *et al.* use a maximum neighbourhood that we will interpret as a dilation using a square structuring element.

$$C1_{b_{s',o}} = C1_{a_{s',o}} \oplus E_{s'} \tag{2.10}$$

where $E_{s'}$ is a square structuring element of a certain size depending on band $b_{s'}$. After this local maximum, a sub-sampling[2] is done using a cell grid and taking the maximum of each cell as the representative for the sub-sampled image. Formally:

$$C1_{s',o}(x_i, y_j) = \max_{x,y \in I(x'_i, y'_j)} (C1_{b_{s',o}}(x, y)) \tag{2.11}$$

where (x_i, y_j) are new pixel indexes that belong to the sub-sampled image, and $I(x'_i, y'_j)$ is the neighbourhood of pixels (x'_i, y'_j) which corresponds to the centre indexes of the cell grid (i, j) on the $C1_{b_{s',o}}$ image.

L5. This layer measures the similarity between the image and the prototypes of the vocabulary (see Universal Vocabulary paragraph below): said layer computes a Gaussian euclidean distance between all prototypes and all possible crops of similar size of the image, once a vocabulary is defined. In this step, they previously change their representation: each image information channel gathers the different scale responses fixing an orientation by just concatenating them:

$$S2_{a_{s'}} = [C1_{s',o_1}, \ldots, C1_{s',o_n}] \tag{2.12}$$

where n indicates that the method considers n different orientations. This layer corresponds to their *S2 units*, which they defined as:

$$S2_{V_i, P_{i,j}, s'} = exp(-\beta \| \sum_o (S2_{a_{s'}} - P_{i,j}) \|^2) \tag{2.13}$$

where $P_{i,j}$ is the prototype j of a determined vocabulary V_i, and β is the sharpness parameter. Notice that *S2 units* are expressed in terms of a certain vocabulary V_i with its prototype $P_{i,j}$ and a specific scale band s' instead of the pair s' and o like previous layers.

To end with their new coding responses, HMAX framework builds *C2 units*, which are result of computing a global maximum across the scale bands and image positions. Hence, only preserve the best match between the prototypes and the image in a vector:

$$\begin{aligned} \overrightarrow{C2} \quad = \quad & [\max_{s'}(\max_{x,y}(S2_{V_1, P_{1,1}, s'}(x, y))), \ldots, \\ & \max_{s'}(\max_{x,y}(S2_{V_1, P_{1,M}, s'}))), \ldots, \\ & \max_{s'}(\max_{x,y}(S2_{V_N, P_{N,M}, s'}(x, y)))] \end{aligned} \tag{2.14}$$

where (x, y) corresponds to the image indexes, M, the size of the vocabulary (the amount of prototypes per vocabulary), s', the scale bands, and N the different vocabularies that have been learned. We consider this final step as a Bag of Words (BOW) representation because vector $\overrightarrow{C2}$ contains NM

[2] Do not confuse this sub-sampling with the down-sampling used in the wavelet transform. This sub-sampling does not follow a power of two reshaping. The neighbourhood of the cell grid depends on the band ([8])

values indexing to a visual word ($P_{i,j}$) or prototype.

Universal vocabulary. T. Serre *et al.* [8] justify the use of a universal vocabulary ($V = V_1 \cup \ldots \cup V_N$) with the performance of their experiments: using a universal vocabulary they achieve good results using a smaller training set compared to those obtained with a specific vocabulary. Below is an explanation of building the universal vocabulary.
Visual words (*prototypes* $P_{i,j}$) are learnt as subset of patches of different sizes in random positions throughout all orientations at $C1_{s'_1,o}$ image description. The amount of different sizes determines the quantity of different vocabularies to be built. For each of the M random images (m_i) selected to learn, they build M prototypes:

$$\begin{array}{rcl} P_{1,m_i} & = & [C1_{s'_1,o_1}(I_1(x,y)), \ldots, C1_{s'_1,o_n}(I_1(x,y))] \\ P_{2,m_i} & = & [C1_{s'_1,o_1}(I_2(x,y)), \ldots, C1_{s'_1,o_n}(I_2(x,y))] \\ \vdots & & \\ P_{N,m_i} & = & [C1_{s'_1,o_1}(I_N(x,y)), \ldots, C1_{s'_1,o_n}(I_N(x,y))] \end{array} \tag{2.15}$$

considering n different orientations, and I_N the neighbourhood of the random positions (x, y) whose size is determined by the specific vocabulary (V_i, $\forall i = 1, \ldots, N$). Accordingly, we could say that they learn N different vocabularies, one for each certain size. In [8] they consider four different orientations and also four different vocabularies of sizes $4, 8, 12, 16$.

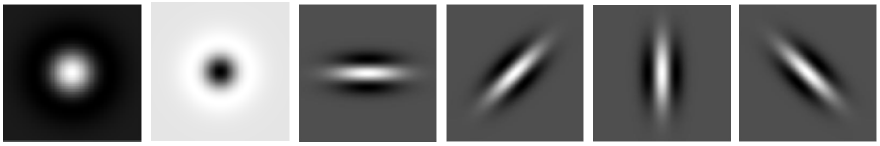
2.4 Malik-Perona Model's

Following the analysis of several bio-inspired models, we focus on the model of J.Malik and P. Perona [14] (MP) based on Julesz' theory [21], since it was finally applied for the texture segmentation. This model can be explained in three main steps that are schematically shown in figure 2.1(b). Like HMAX, the MP method works on gray-scale images.

L1. In this first stage, each gray-scale image (I) is convolved with a set of linear filters ($F_{s,o}$) to simulate the function of simple cells and obtain the output of V1. To make this convolution, they chose Differences of Offset Gaussians (DOOG - equation 2.16) filters [22]:

$$\begin{array}{lcll} DOG1(\sigma) & = & aG(0,0,\sigma_i,\sigma_i) - aG(0,0,\sigma_0,\sigma_0) & [\sigma_i : \sigma : \sigma_0 = 0.71 : 1 : 1.14] \\ DOG2(\sigma) & = & -aG(0,0,\sigma_i,\sigma_i) + 2aG(0,0,\sigma,\sigma) - aG(0,0,\sigma_0,\sigma_0) & [\sigma_i : \sigma : \sigma_0 = 0.62 : 1 : 1.6] \\ DOOG2(\sigma) & = & -aG(0,\sigma,\sigma_x,\sigma_y) + 2aG(0,0,\sigma_x,\sigma_y) - aG(0,-\sigma,\sigma_x,\sigma_y) & [\sigma_x : \sigma : \sigma_y = 3 : 1 : 1] \end{array} \tag{2.16}$$

where $G(x_0, y_0, \sigma_x, \sigma_y)(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp -((\frac{x-x_0}{\sigma_x})^2 + (\frac{y-y_0}{\sigma_y})^2)$ is an ordinary 2-D Gaussian function with a certain displacement with different standard deviations for the axis x and y (see figure 2.2).



(a) (b) (c) (d) (e) (f)

Figure 2.2: Set of DOOG filters. (a) DOG1, (b) DOG2, (c,d,e,f) set of different oriented DOOG filters.

L2. As is known, there are some non-linearities in human receptive fields. Following this idea, MP used half-wave rectification. Their choice is justified, since it does not lose the sign of filter response [14]. Half-wave rectification consists in separating positive and negative part of the filtered image. In short, in this layer two different responses are obtained:

$$\begin{aligned} R_{i_{s,o}} &= (I * F_{s,o})^{+} = max\{(I * F_{s,o}), 0\} \\ R_{i+1_{s,o}} &= (I * F_{s,o})^{-} = max\{-(I * F_{s,o}), 0\} \end{aligned} \tag{2.17}$$

L3. Malik and Perona realized that a second non-linearity function was needed. The use of a half-wave rectification did not allow discrimination of texture pairs composed of opposite patterns [8]. To solve this problem, they defined a threshold:

$$T_{i_{s,o}}(x,y) = \max_{j_{s,o}} \max_{x,y \in I_{j_{s,o} i_{s,o}}(x_0,y_0)} \alpha_{j_{s,o} i_{s,o}} R_{j_{s,o}}(x,y) \tag{2.18}$$

where $I_{j_{s,o} i_{s,o}}$ corresponds to the neighbourhood of (x_0, y_0) in which channel $j_{s,o}$ inhibits neurons in channel $i_{s,o}$, and $\alpha_{j_{s,o} i_{s,o}}$ is a measure of the effectiveness of this inhibition. Using this threshold, they define the post-inhibition response for channel $(i_{s,o})$, which selects high responses:

$$PIR_{i_{s,o}}(x_0,y_0) = \max_{x,y \in S_{i_{s,o}}(x_0,y_0)} [R_{i_{s,o}}(x,y) - T_{i_{s,o}}(x,y)]^{+} \tag{2.19}$$

where $S_i(x_0, y_0)$ is a sub-sampling neighbourhood of (x_0, y_0). Then, the set of PIR_i become the final outputs for the MP method.

Chapter 3

Towards a more accurate low-level method

After the review of some low-level methods, we focus our analysis to the ID models. In this chapter we point out how ID has been applied to predict colour induction or saliency ([5], [4], [3], [6]) in section 3.1. Taking into account that ID has presented interesting properties in different visual tasks, we explore its stages in section 3.2 which entail to investigate new designs for each stage pursuing the saliency prediction visual task. This chapter ends with section 3.3 where we bring closer to the concept of visual codes, followed by section 3.4 where are summarised all the alternatives proposed.

3.1 Induction-derived family models

Chapter 2 presents the main stages of ID family models. Although the motivation of this model was the brightness assimilation and contrast effects, it was also applied to saliency estimation. Therefore, it gives rise to different methods and this is the reason why we call Induction-Derived family to the set of algorithms that share their first stages. In this section we show how was adapted to predict colour induction and visual saliency.

3.1.1 Predicting colour induction

Colour induction is a well-known phenomena that has been studied for long. It refers to the effect of perceiving a different colour than its own real colour in a certain patch due to the presence of other coloured patches on its surrounding. Figure 3.1 shows this effect, where each couple of circles have a central ring with the same colour, in spite of our perception is different.

ID model was applied in order to predict this effect. First, it was used to predict brightness induction (grey-scale) in BIWaM [3] and after, in CIWaM method [4] for colour induction. The difference between

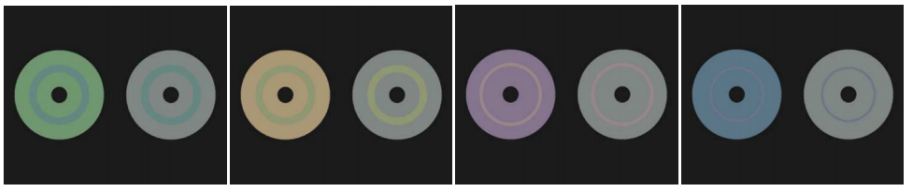


Figure 3.1: (Image from [4]). Example of colour induction effect. Each couple of circles have a central ring with the same colour although we perceive them different due to the colour of the circle.

both methods falls on the ECSF applied. These methods were able to recover the perceived image following next equation:

$$I_{perceived} = \sum_{s=1}^{n} \sum_{o=v,h,d} \alpha_{z_{s,o}} \cdot w_{s,o} + c_n \tag{3.1}$$

where c_n is the residual of the wavelet transform and n, the spatial frequencies. The rest of the functions are derived from the equations in section 2.2. ID model could be applied to solve the prediction of colour induction effects thanks to the fact of including properties as spatial frequency channels, orientation of receptive fields and the centre-surround energy measure, which are relevant for the colour induction problem.

3.1.2 Predicting visual saliency

The evolution of ID was extended to solve the problem of saliency estimation. This visual task refers to the distinction of some object with some special property (or properties) that stands out this object from their neighbours and causes the attraction of our attention. As the centre-surround energy measure becomes an important property to detect saliency regions in an image, N. Murray *et al.* in [5] (SIM) and [6] (SIM + GT) extended ID to predict saliency estimation. SIM + GT is an extension of SIM, which adds geometrical grouplets to improve SIM method. In these models, ECSF was slightly modified compared to the one used in [4]: a least square regression was run to adjust ECSF parameters (see [5] for review). To build saliency maps, N. Murray *et al.* propose an inverse wavelet transform of the α weight for each channel. Finally, in order to combine all channels information and get a single saliency map, they compute the Euclidean norm.

$$S_c(x,y) = WT^{-1}(\alpha_{z_{s,o}}(x,y)) \tag{3.2}$$

$$S = \sqrt{S_1^2 + S_2^2 + S_3^2} \tag{3.3}$$

again, $\alpha_{z_{s,o}}$ is defined in section 2.2.

3.2 Exploring the ID model stages

We have seen that ID presents properties generalizing different visual tasks. It lead us to explore better the method and to try to improve the model (which were initially fitted on colour induction data) using saliency datasets. Our analysis is done under a global hypothesis that requires to be tested by a layer analysis:

Global hypothesis: The elements selected to implement the layers of the model can constraint its capabilities to define more generic visual codes.

The layer analysis entails to other hypothesis that we try to prove in following subsections. In an introductory way, we sum up our layer analysis as the answer to next questions:

L1 : Is wavelet decomposition reducing representation capabilities?

L2: Centre-surround regions computed on the wavelet decomposition are constrained to a fixed rectangular. Is there any shape of centre-surround that can improve the method?

L3: Is the quadratic rational function limiting the divisive normalisation step?

L4: Is ECSF relevant for the method? Can it be improved?

3.2.1 Can feature-shape selectivity be improved?

SIM model takes advantage of mathematical wavelet properties. The wavelet transform is obtained using a family of 1-D Gabor Filters. Analysing the method, we notice that wavelet transform tend to detect object's edges of the image.

We believe that saliency is more related to blob detection than edge detection. In fact, human attention derives from an enhanced object with respect to its surroundings. In this manner, we propose using the family of DOOG filters, which contain both blob and bar detectors. This change should be able to get different shapes responses related to circles or ellipses. To accelerate computational time in these complex operations, we work in Fourier domain.

3.2.2 Does centre-surround window shape matters?

Following our idea of relating saliency to blob detection, we propose to change centre-surround regions. Centre-surround regions determine the area where both energy are compared. For this reason, it is an important point in the method which has to be in keeping with the set of filters used in linear filtering stage. We could think of centre-surround responses as the set of stimuli in our human visual system, each stimulus being related to a specific shape. SIM method uses constraint rectangles to consider the centre-surround regions. In order to be able to answer the question of this subsection, we propose as adapted centre-surround shapes a set of ellipses of different orientations and sizes (adapted to DOOG filters used in first stage). These change corresponds to the pattern detection (oriented ellipses) that are in V1 (see figure 1.2), which should be able to detect more complex shapes.

3.2.3 Does divisive normalisation function affect?

We have indicated previously that cell responses tend to saturate high contrasts. Although SIM method tends to predict saliency correctly, we notice that saliency predictions were usually noise images. We attribute this fault to the divisive normalisation (L2) that SIM uses, the quadratic rational function, (see equation 2.3) due to its capability of saturating energy contrast for small values. Since physiological studies relate the saturation of high contrasts to functions with a sigmoid shape [16], we propose using a parametric sigmoid function in this step:

$$z_{x,y} = f(r_{x,y}) = \frac{1}{1 + e^{-\alpha \cdot (r_{x,y}) + \beta}} \tag{3.4}$$

Our idea is, then, to saturate contrast with higher values than those which are saturated by the quadratic rational function.

3.2.4 Is ECSF relevant for ID?

ID family models uses ECSF with slightly changes in ECSF parameters. We want to analyse its effect in the model and its relevance. Regarding ECSF's goal, which is to enhance or discard the responses of L3, our idea is to discover if it is needed or not. If L3 responses are accurately detected the effect might not be needed. For this reason, we will attempt to use a Null-ECSF, which corresponds to consider $\alpha_{s,o} = z_{s,o}$.

3.3 Towards visual codes and learning parameters

The term of visual codes is a new trend in computer vision and they represent step forward in human visual system in terms of image representation. One of their properties is the loss of the retinal composition. In this work we try to bring SIM closer to this concept. Next sections explain our proposal to change the representation of the image on SIM pipeline which also allow to learn ECSF appropriately.

3.3.1 Towards visual codes

In computer vision research, several algorithms have been developed to face the problems of object detection, description, recognition. Most bio-inspired models consider neuron-responses when defining their feature space. Object features are described as a response vector in a high dimensional space. We can relate the dimension of this space to the amount of neurons that have been considered [7]. We propose a new feature representation to get it closer to this response vector.

SIM has a weighting function which became a particular point in the model, considering that is the only existing algorithm that includes a weighting function. This weighting function (ECSF) is defined in the space centre-surround energy and scale. In this way, we propose a feature vector of a set of visual words which are described from centre-surround energy and scale. The following lines describe our proposal.

Let $Z-S$ space be the Centre-Surround Energy and Scale Space. We quantify Z axis in Z_n bins, while S axis in S_n bins, which divide our $Z-S$ space in $B_{Z_n \times S_n}$ bins. We define δ-function as follows:

$$\delta_m(z_{o,s,c}) = \begin{cases} 1 \text{ if } z_{o,s,c}(x,y) \in B_m \\ 0 \text{ OTHERWISE} \end{cases} \tag{3.5}$$

where m indicates the index of Bin m of our quantification ($m \in [1, Z_N \times S_N]$). Using this δ-function we can describe each pixel by $\overrightarrow{v}$ vector, which is a set of $\delta_{Z_n \times S_n}$, for each orientation, scale and color-channel (c):

$$\overrightarrow{v}_{o,s,c}(z_{s,o}) := (\delta_1(z_{o,s,c}), \ldots, \delta_{Z_n \times S_n}(z_{o,s,c})) \tag{3.6}$$

Notice that our new representation does not follow a pyramid multi-resolution as original method had applied. Each scale is represented by the same amount of pixels. However, wavelet transformation could be done without taking advantage of the multi-resolution decomposition.
Here, we have to incorporate the weighting function. For this, we quantify the function in $B_{Z_n \times S_n}$ bins, as before. We will have $A_1 \ldots A_{Z_n \times S_n}$ different weights to apply. We define another vector $\overrightarrow{A} = (A_1, \ldots, A_{Z_n \times S_n})$ which will weight our $\overrightarrow{v}$.

To build the saliency map S_c, instead of using the inverse wavelet transform and an Euclidean norm (2-norm) we simplify it using the 1-norm or taxicab norm ($\|x\| = (\sum \|x_i\|^p)^{(1/p)}$, $p = 1$), adding up all pixel information for the scales, orientations and color-channel responses:

$$S(x,y) := \sum_c \sum_s \sum_o \alpha_{o,s,c}(x,y) \tag{3.7}$$

where $\alpha_{o,s,c}(x,y) = \sum_{i=1 \ldots Z_n \times S_n} A_i v_i$.
Although we do not end with the building visual codes because we are still keeping location information, but we could use our new representation to get them just counting how many times each δ_i appears in the whole image.

3.3.2 Learning the weighting function

ID has a particular layer (L4) which modifies centre-surround contrast energy. We have seen that BIWaM, CIWaM and SIM uses this weighting function slightly modified between them. In fact, we believe that ECSF can be customised according to the problem to be solved. Considering our method of building saliency maps, we are able to learn the weighting function (A_i) by a simple linear SVM that can learn the adapted ECSF to the visual task. In this line, each component of the normal vector to the hyperplane that finds SVM will correspond to the set of weights (A_i). Remember that SVM classification is based on the sign of the decision function ($f = \sum_{s,c,o}(\overrightarrow{A}\,\overrightarrow{v}_{o,s,c}) + b$). Besides, SIM adapts a function which was fitted on colour induction data that may be improved and can improve the model.

3.4 Synopsis

In this section we sum up the different alternatives for ID that we have proposed in this dissertation. In figure 3.2 we break down the ID framework in its main steps with the different methods that we are studying. Due to it is pursuing an accurate model from SIM, we refer to it as Accurate Saliency by Induction Mechanisms our new proposal (ASIM). The different options are summarized in the following lines:

L1. In image filtering step, Gabor multi-resolution (GM) and Gabor (G) relate to the use of 1-D family Gabor filters on a multi-resolution Gabor wavelet transformation and Gabor wavelet transformation, respectively; while DOOG (DG) is related to the use of DOOG family filters.

L2. In centre-surround step we talked about unshaped centre-surround (UCS) when the centre-surround responses are calculated in oriented and constrained rectangles, while shaped centre-surround refers (SCS) to elliptical areas. Gabor multi-resolution and Gabor are run using unshaped centre-surround, while DOOG uses shaped regions.

L3. For divisive normalisation we compare the use of quadratic rational function (QRF) with the use of a parametric sigmoid function (SF).

L4. We study different ways to apply a weighting function: continuous ECSF (CW), which refers to the original ECSF applied in [5]; quantified ECSF (QW), which is just the original ECSF quantified in a certain amount of bin or SVM-ECSF (SW) when the ECSF is learned by SVM technique. We also consider a Null-ECSF (NW), which does not apply any weight to the centre-surround energy measure.

Recovery To recover saliency maps, the method of N. Murray *et al.* uses the inverse wavelet transform followed by an Euclidean Norm . We will call this way as Euclidean saliency construction (EN). The 1-norm (1N) construction will refer to our assumption of building saliency maps just adding up channels, scale and orientation responses.

Using our nomenclature, SIM method [5] corresponds to the use of GW + UCS+QRF + CW + EN.
In this paper we propose a new feature representation. It is an intermediate step between the divisive normalisation and the induction weights, which is only compatible with responses that preserve the sizes of the original image. Therefore, it can only be applied when image filtering is done by Gabor or DOOG.

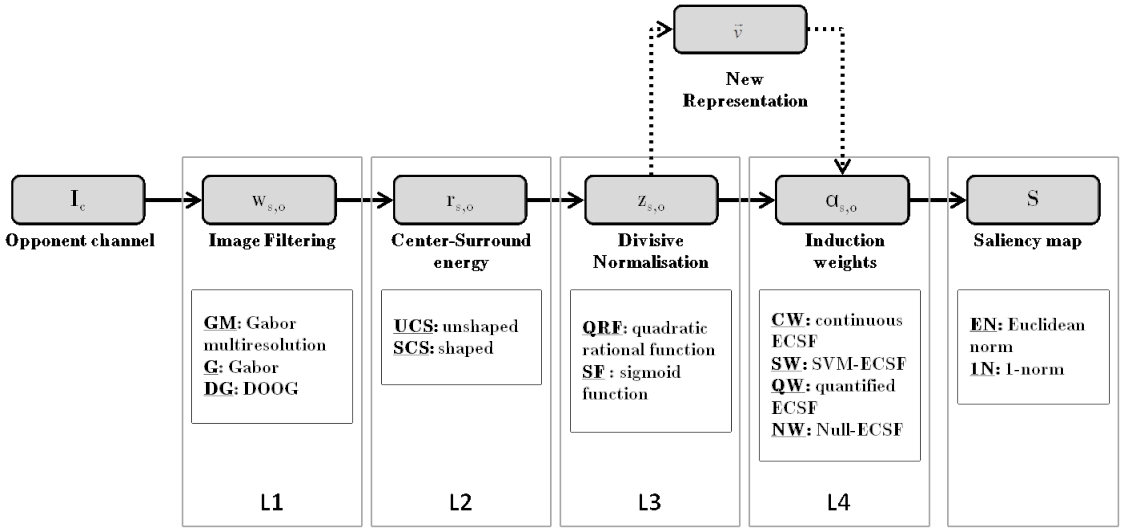


Figure 3.2: Several alternatives for each layer on SIM method.

Chapter 4

Experiments and Discussion

In this chapter we are going to perform some experiments in order to be able to get an answer for the open questions of chapter 3 by working on the problem of saliency estimation. We also discuss how the results of mentioned experiments are affected by each change in the pipeline and it attempts to understand the roots of the ID. First, in section 4.1 we discuss the metrics that usually are used when evaluating saliency estimation performance. Also, we propose a new metric which takes into account the noisy-effect of the estimations. Next, we present the database (section 4.2) where we run the experiments (section 4.3).

4.1 On measuring Saliency Map Performance

Most saliency studies are evaluated by Kullback-Leibler divergence (KL) and area under ROC curve (AROC) measurements [23]. The usual procedure concerning the use of these measurements is to compare two different histograms. On the one hand, histograms at fixation points (true positives) are built taking saliency values at these determined locations. On the other, histograms at non-fixated points (false positives) are computed using fixated-points of other random images. Although some fixation points have low saliency values on fixation map[1] (less relevant points), they do not take into account this importance rate. KL divergence measures the capability of differentiating both histograms using following equation:

$$KL(P,Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i) \tag{4.1}$$

being P and Q two discrete probability distributions (histograms of saliency values), with $Q(i) > 0 \ \forall i$. Usually, Q corresponds to the histogram of saliency at fixation points, while P, at non-fixation points.

[1] The term fixation map refers to the saliency map built from obtained fixation points. They follow a Gaussian distribution around fixation points where subjects pay more attention (time of observation). They enhance more relevant fixation points with a higher value in the fixation map.

Using KL metric, what is considered to be a good saliency estimation occurs when both histograms are different enough. But we think that there is not enough evidence to assume false positives to be non-fixation points, since most of saliency objects are in the centre of the images. Nevertheless, they also use AROC measurement. ROC curve is defined varying the threshold of saliency value of true positive and false positive sets. It is, therefore, a measurement of the ratio of saliency values between true positive rate and false positive rate. This measure allows to analyse the differences between two sets on their corresponding saliency values.
After several experiments we notice that noisy saliency maps have an advantage over the others using these metrics. This effect could be a result of only taking into account information of few pixels. Therefore, we question these measurements due to:

- considering false positives as fixation points from other images expecting to have low saliency values.
- considering few pixel information to evaluate saliency map
- the ignorance of the fixation maps (some fixation points are more relevant than others)
- the favouritism for noisy images.

To avoid these problems, we propose a new metric based on a new definition of a good saliency prediction. We interpret a good saliency estimation to be those saliency maps whose fixation points have enough high values in saliency predictions and also, the amount of pixels with these high values represents a low percentage of covered area in the image. Notice that with this definition we make the assumption that saliency maps follow a Gaussian-distribution around each fixation point. This percentage of area is a measurement that allows us to detect noisy saliency predictions.

$$\text{True Positives Rate} = \frac{\#\{f_p | I_{SMap}(f_p) > th\}}{\#\{f_p\}} \tag{4.2}$$

$$\text{Percentage of saliency area} = \frac{\#\{x_i | I_{SMap}(x_i) > th\}}{\#\{x_i\}} \tag{4.3}$$

where f_p corresponds to the fixation points, I_{SMap} to the saliency estimation map, th the threshold, and $\{x_i\}$ the set of pixels of the image, being $\#\{x_i\}$ the number of pixels of the saliency estimation map (area). Therefore, we can relate our true positives rate to the recall and the percentage of saliency area to the precision measurements. Following the metric used in [24], we use a new Recall-Precision curve. Once again, the problem is interpreted as a binary classifier, where varying the threshold (saliency value) we can build the Recall-Precision curve. The curve measures the relation between true positives and percent of saliency area. In this line, we define true positives as those fixation points that have higher values than threshold on saliency estimation image. In our metric we emphasise those fixation points with high values, since we believe that it is more important to detect a fixation point with higher value in a fixation map than lower ones. To this end, we replicate each fixation point n times depending on their saliency values in fixation maps.

We will call this new metric Weighted Area under Recall Precision curve (WARP). Like the other two metrics, they are directly proportional to the efficiency of the method: higher values indicate a greater performance.

4.2 The dataset

The experiments are run in Bruce & Tsotsos [25] dataset which contains 120 color images of 511x681 pixels. The dataset also provides their eye-fixation obtained from 20 different subjects and the corresponding fixation maps. We will use the metrics explained in section 4.1: KL, AROC and WARP to see how they are affected by the changes. To neutralise random effects in KL and AROC, these metrics are computed 100 times and we are also able to get the standard error. In case of WARP we also show the standard error between all the different results obtained for each one of the 120 images.

4.3 Results & analysis

Each following subsection discusses a specific part of the pipeline which uses different options of our proposal (see figure 3.2). We refer to original SIM when using the Gabor multi-resolution family filters (GM), unshaped centre-surround regions (UCS), quadratic rational function (QRF) in divisive normalisation step and continuous ECSF (CW). ASIM will refer when some changes are applied respect to original SIM. As both reconstructions proposals achieve the same results (see subsection 4.3.1), we will use 1-norm in most of experiments, although original SIM uses the Euclidean norm.
First of all, in subsection 4.3.1 we will analyse the effect of simplifying the reconstruction step in saliency prediction. Next, we will start with the family of filters effects in subsection 4.3.2. Several divisive normalisation function are evaluated in subsection 4.3.3, followed by the study of the importance of the weighting function in subsection 4.3.4. Subsection 4.3.5 analyses the effect of our new representation and its quantification, including some results learning ECSF by SVM. Finally, in subsection 4.3.6 we show a comparison of the contribution of main changes.

4.3.1 Analysis 1: Recovery function

In this subsection we analyse the effect of the small change proposed in the reconstruction step. We want to guarantee if the results in euclidean norm and 1-norm are similar. We run original SIM using both ways of recovering saliency maps. Table 4.1 shows that there are practically no differences. Due to its simplification, feature experiments will be done using the 1-norm.

4.3.2 Analysis 2: Family of filters

In this experiment, we analyse the influence of the family of filters used in L1. Table 4.2 shows our performance using both ways of using a family of 1-D Gabor filters and DOOG. Practically all methods

Reconstruction	KL (SE)	AROC (SE)	WARP (SE)
Euclidean	0.4265 (0.0030)	0.7013 (0.0008)	0.7948 (0.0947)
1-norm	0.4348(0.0031)	0.7011 (0.0007)	0.7961 (0.09509)

Table 4.1: Performance of different ways to reconstruct Saliency Maps.

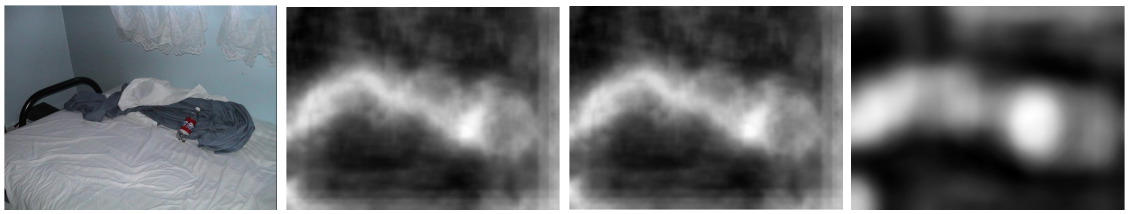
achieve the same results, the ones which use Gabor filters performing better attending KL and AROC measures. Nevertheless, DOOG obtains better results in WARP measure. Also, saliency predictions obtained with DOOG are qualitatively cleaner with respect to the others (see figure 4.1). All family filters require a set of parameters that had to be adjusted: number of scales, sizes and orientations. Regarding to filter size, DOOG filter were defined resembling the sizes of Gabor filters used in [5]. We compute the pipeline using six different scales and three orientations for Gabor filters and five for DOOG (isotropic, 0, 90, 45 and 135 degrees). Bearing in mind the results, we conclude that our new filtering stage is consistent with the ID, achieving a better performance in terms of WARP.
With this experiment, we are able to answer the questions of subsection 3.2.1 and 3.2.2. We answer both together due to constraint relation of the centre-surround shape with the type of linear filtering. Although quantitative results do not present significant differences, qualitatively results shows that saliency estimations are different depending on the method used in L1. In fact, DOOG gets more accurate estimations. Our new proposal using DOOG and shaped centre surrounds improve the selectivity of image regions due to its tend of selecting blob-enhanced regions in the image with more orientations and shapes and also allow to detect more complex shapes.

Family of filters	KL (SE)	AROC (SE)	WARP (SE)
Gabor multi-resolution	0.4348(0.0031)	0.7011 (0.0007)	0.7961 (0.09509)
Gabor	0.4301 (0.0031)	0.7011 (0.0007)	0.7928 (0.0924)
DOOG	0.4184 (0.0033)	0.6925 (0.0007)	0.8119 (0.0087)

Table 4.2: Performance of the method using different filtering decomposition

4.3.3 Analysis 3: Divisive normalisation

In order to increase or decrease the saturation of high responses, we develop a study of several values for α and β on parametric sigmoid function (equation 3.4) for the original SIM. The results are shown in table 4.3 where we can notice that the best values are achieved using the ordinary sigmoid function (for $\alpha = 1$ and $\beta = 0$) and setting $\alpha = 2$ and $\beta = 1.5$. Notice that we only show KL performance, since AROC is practically the same in all cases (0.6915 ± 0.0056 (0.0007)). As we can see in table 4.3, there is not much difference between the values. The effect of sigmoid is minuscule comparing with the



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

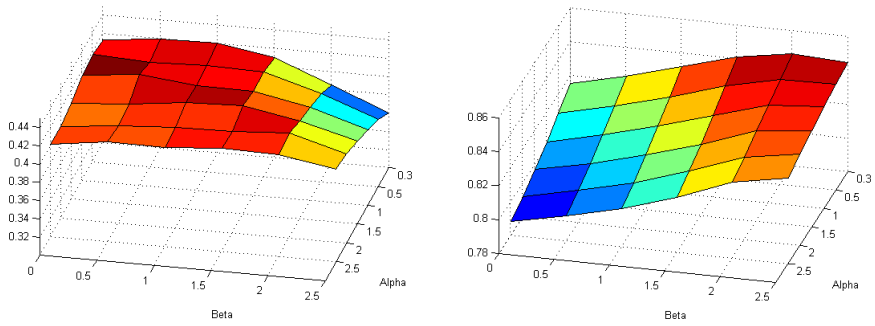
Figure 4.1: The effect of the type of filters used in L1 on saliency predictions. (a) Original image, and saliency predictions of ID using (b) GM, (c) G and (d) DG.

rational quadratic divisive normalisation (which obtains a KL of 0.4348), but it allows better results. With this experiment we notice that higher values correspond to functions that saturate smaller values than the quadratic rational function.

α \ β		0	0.5	1	1.5	2	2.5
0.3	KL	0.4087	0.3983	0.3984	0.3761	0.3709	0.3594
	(SE)	(0.0032)	(0.0031)	(0.0031)	(0.0024)	(0.0025)	(0.0027)
0.5	KL	0.4148	0.4202	0.4052	0.3973	0.3906	0.3730
	(SE)	(0.0032)	(0.0028)	(0.0030)	(0.0029)	(0.0030)	(0.0028)
1	KL	**0.4401**	0.4296	0.4304	0.4256	0.4179	0.4098
	(SE)	(0.0034)	(0.0029)	(0.0031)	(0.0031)	(0.0032)	(0.0030)
1.5	KL	0.4320	0.4377	0.4350	0.4375	0.4287	0.4279
	(SE)	(0.0026)	(0.0031)	(0.0035)	(0.0032)	(0.0031)	(0.0028)
2	KL	0.4289	0.4323	0.4307	**0.4436**	0.4323	0.4296
	(SE)	(0.0028)	(0.0032)	(0.0036)	(0.0031)	(0.0033)	(0.0033)
2.5	KL	0.4237	0.4162	0.4187	0.4381	0.4346	0.4350
	(SE)	(0.0033)	(0.0034)	(0.0029)	(0.0032)	(0.0033)	(0.0034)

Table 4.3: Divisive Normalisation using sigmoid function performance in function of scale (α) and translation (β) parameters.

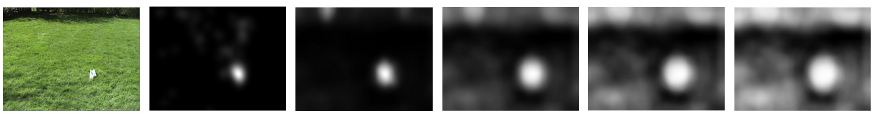
We develop the same experiment using DOOG filters in L1. We show the results in terms of KL and WARP in figure 4.2. As before, AROC achieves the same results in all cases (0.6890 ± 0.0047 (0.0007)). We enhance the case of sigmoid of $\alpha = 0.5$ and $\beta = 2.5$, where we achieve 0.84635 in WARP. KL behaviour obtained is similar to the previous experiment (see table 4.3): higher values are achieved



(a) KL measurement (b) WARP measurement

Figure 4.2: Parametric sigmoid performance in KL and WARP measurement varying α and β parameters.

in small β values. Nevertheless, WARP behaviour is opposite to KL, since saliency estimations are qualified better using WARP for higher β values. We attribute the effect of the two opposite behaviours in both measurements to the trend of KL of having advantage in noisy images. Regarding to WARP, the performance is better when the saturation is done for higher values than quadratic rational function, following our intuition of changing the divisive normalisation formula.



(a) (b) (c) (d) (e) (f)

Figure 4.3: Different saliency predictions obtained for the original image (a) with whose fixation map is (b). Analysing from left to right, KL measurement goes from less performance to high performance, while WARP goes from high performance to less. Therefore, saliency prediction (c) achieves the best performance in WARP, but the worst using KL. Different to saliency prediction (f), which achieves the best performance for KL but the worst for WARP.

Figure 4.3 shows different saliency predictions obtained for a specific image. We can observe that noise-effect incorrectly affects KL performance, while WARP follows human analysis to assess saliency estimation. Besides, qualitatively saliency estimations change significantly and measurements indicate that divisive normalisation function affects to the model, solving the question of subsection 3.2.3.

Figure 4.4 shows the different parametric sigmoid function that have been tested. It enhances the two best curves using KL measurement (in red) and WARP (in green) measurement. If we compare them to the quadratic rational function, green lines saturate few responses and, therefore, follows our intuition of reducing noise.

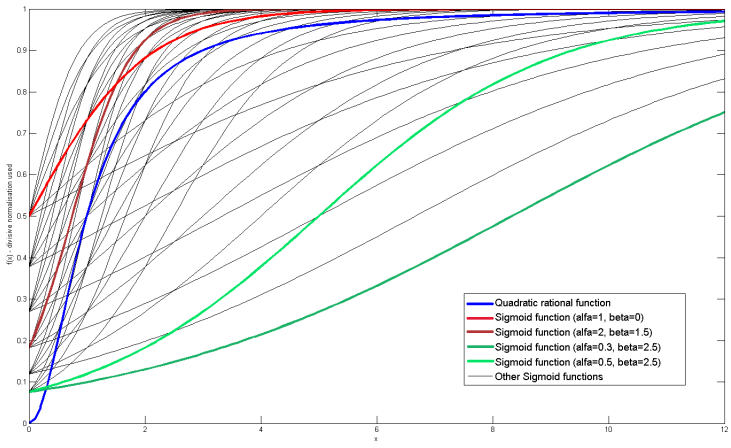


Figure 4.4: Set of parametric sigmoid function tested. In red, the best performance obtained in KL terms. In green, the best performance in WARP. Blue line corresponds to the quadratic rational function.

4.3.4 Analysis 4: Weighting function or ECSF

In this experiment we want to analyse the ECSF's effect and try to answer the question of subsection 3.2.4 related to the relevancy of the ECSF. We run the original SIM using different types of ECSF and also using ID with DOOG filters and the different types of ECSF. In table 4.4 we show our quantitative results for both families of filters. The continuous ECSF achieves the best quantitative results, followed by a Null-ECSF. ID incorporates this weighting function to enhance or discard specific responses. An important conclusion of this experiment is that its contribution is more emphasised in Gabor Multi-resolution than DOOG. For this reason, when adding a Null-ECSF to a Gabor multi-resolution the performance decreases significantly while in DOOG, the slope is practically unnoticeable. This effect can be justified due to our accurate way of defining the centre-surround energy measure.
In figure 4.5 we show the prediction saliency images for a specific image. In this figure, we can qualitatively conclude the same performance as table 4.4. There is an important factor that is present in this figure: Null-ECSF with Gabor multi-resolution shows the trend of centre-surround energy measure to be the edges of objects in the images, since Null-ECSF implies considering only centre-surround energy

measures as responses. This edge-selection is reduced in DOOG pipeline.

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

Figure 4.5: The effect of the weighting function type. (a) Original image, (b) GM + CW, (c) DG + CW, (d) GM + NW , (e) DG + NW.

Family of Filters	weighting function	KL (SE)	AROC (SE)	WARP (SE)
Gabor multi-resolution	Continuous ECSF	0.4348(0.0031)	0.7011 (0.0007)	0.7961 (0.09509)
	Null-ECSF	0.3396 (0.0023)	0.6781 (0.0006)	0.7779 (0.0859)
DOOG	Continuous ECSF	0.4184 (0.0033)	0.6925 (0.0007)	0.8119 (0.0087)
	Null-ECSF	0.4029 (0.0029)	0.6946 (0.0007)	0.8053 (0.0857)

Table 4.4: Saliency Predictions results using different types of ECSF in original SIM and using DOOG filters in ID pipeline.

4.3.5 Analysis 5: Coding outputs and learning parameters

In this case, we want to prove that our new representation does not affect our pipeline. For this reason, we use original SIM configuration for each step but using Gabor and we add our representation. Due to this new representation, in order to get corresponding induction weights we quantified the original ECSF from Otazu *et al.* in n bins, where n is the total number of bins used to build our new vector $\overrightarrow{v}$. Also, we analyse the influence of the quantification of the $Z - S$ space in the final results. In table 4.5 the obtained values are shown. It is clear that our new representation is not counter-productive in our framework comparing with the continuous case. Also, we consider that a quantification of 20 bins is sufficient for our goal.

We have seen that ECSF is a particular function on ID. This ECSF function was defined by Otazu *et al.* [4] to predict colour induction effects on images. They adjusted two Gaussian functions according to results of psychological experiments. Afterwards, N. Murray *et al.* readjusted the two Gaussian functions to the problem of saliency estimation by minimum least square error. Now, we try to learn the weighting function without imposing the shape of two Gaussian functions. Table 4.6 shows the results

n	KL (SE)	AROC (SE)	WARP (SE)
continuous	0.4301 (0.0031)	0.7011 (0.0007)	0.7928 (0.0924)
20	0.4301 (0.0031)	0.70008 (0.00074)	0.7936 (0.0924)
50	0.4336 (0.0033)	0.70068 (0.00074)	0.7932 (0.0924)
70	0.4342 (0.0033)	0.70046 (0.00074)	0.7931 (0.0925)
100	0.4365 (0.0033)	0.70069 (0.0074)	0.7930 (0.0924)

Table 4.5: New representation and its quantification (n bins) effect on final saliency predictions.

we get. To build training and testing sets, we consider as true positives (saliency points) those fixation points with a higher value than a threshold (60) in the fixation map. Using this way, we get the n saliency points to train for an image. We also consider n non-saliency points selecting n locations in the image which does not exceed this threshold. Therefore, we achieve a balanced database. In the case of Gabor in linear filtering step, SVM is not able to learn a function as good as the original one, but learned ECSF in DOOG achieve the same results. We attribute the poor learning in Gabor case to the noisy saliency prediction images that ID gets using Gabor in its Image Filtering. In fact, WARP measurement increase using learned ECSF even in Gabor pipeline.

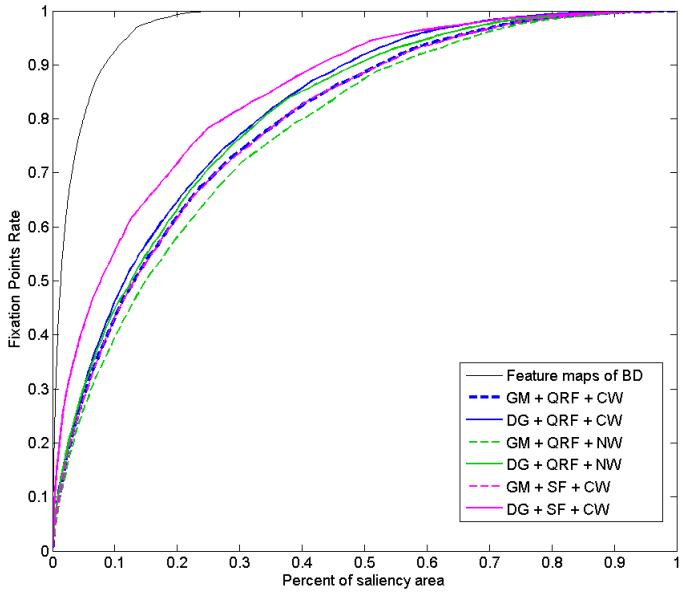
filtering	type ECSF	KL (SE)	AROC (SE)	WARP (SE)
Gabor	SVM-ECSF	0.3579 (0.00277)	0.68152 (0.000688)	0.82309 (0.07994)
Gabor	original ECSF	0.4301 (0.0031)	0.7011 (0.0007)	0.7928 (0.0924)
DOOG	SVM-ECSF	0.4159 (0.00312)	0.69493 (0.00716)	0.8101 (0.08604)
DOOG	original ECSF	0.4184 (0.0033)	0.6925 (0.0007)	0.8119 (0.0087)

Table 4.6: Performance comparison between original ECSF and learned ECSF by SVM.

4.3.6 Global comparison & conclusions

To end with the quantitative results, we show figure 4.7 which relates several experiments in terms of WARP. We gather on it several experiments in order to analyse the effect of each change. Black line corresponds to the evaluation of WARP in fixation maps provided by the database. Continuous lines refer to the use of DOOG in L1, while discontinuous lines to the Gabor wavelet multi-resolution. Original SIM is represented by a discontinuous blue line, which only surpass the case when Gabor multi-resolution is combined with the quadratic rational function with a null ECSF. Besides, we do not get a better WARP performance in any experiment using Gabor multi-resolution. We just achieve a similar performance when divisive normalisation is done by the sigmoid function with parameters $\alpha = 0.5$ and $\beta = 2.5$. Attending the results of DOOG, all the changes outperform the results of Gabor multi-

resolution. When the quadratic rational function is used in divisive normalisation step, original ECSF and null ECSF get similar results, although original ECSF is still better. As important contribution, we stand out the sigmoid function, which outperforms considerably all the other combinations, considering that its Recall-Precision curve is the closest to the one gotten for the fixation maps of the database.



Method	WARP
Fixation maps	0.9713
GM + QRF + CW	0.7961
DG + QRF + CW	0.8119
GM + QRF + NW	0.7779
DG + QRF + NW	0.8053
GM + SF + CW	0.7953
DG + SF + CW	0.8464

Table 4.8: WARP for each stage setting.

Table 4.7: Recall-Precision curves for each stage setting.

To end this chapter, figure 4.6 shows saliency estimations using Original SIM and our ASIM. According to our experiments, ASIM is run using DOOG filters and a sigmoid function in divisive normalisation step with $\alpha = 0.5$ and $\beta = 2.5$. Responses are codified using our vector $\overrightarrow{v}$ in a quantification of 20. It is clear that our new way to use ID predicts better saliency maps, WARP follows human evaluation, and is also closer to the fixation maps.

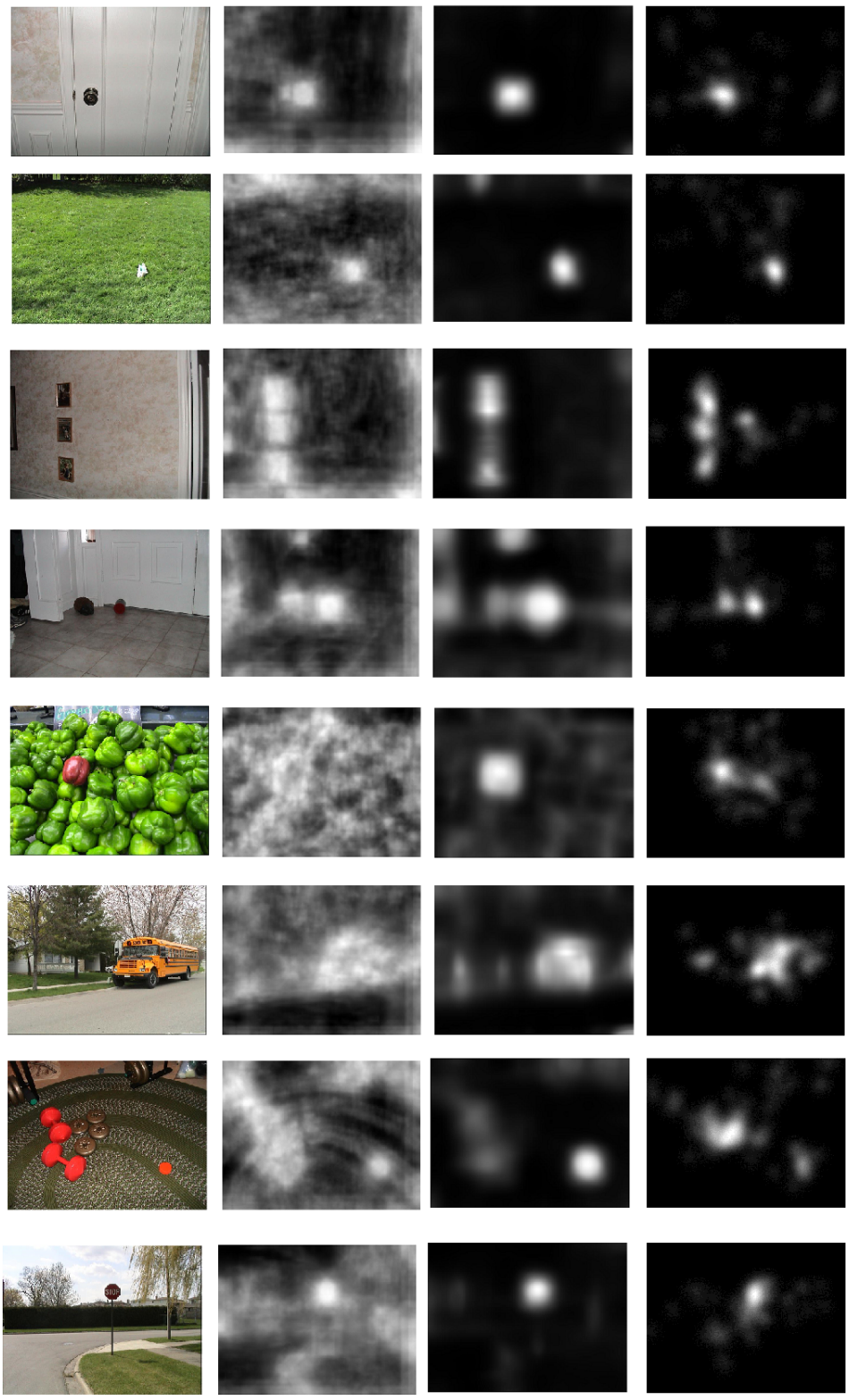


Figure 4.6: Saliency maps Response. For each pack, from left to right, original image, SIM saliency prediction, ASIM saliency prediction and fixation maps from database.

Chapter 5

Conclusions & further work

To end with this dissertation, this chapter shows our conclusions of our exploration and analysis of low-level vision models in section 5.1. Although in saliency estimation our results outperform SIM in new measurement and allow to get closer saliency maps to the fixation maps provided by the database, several parts can still be improved. The new lines proposal of research are listed in 5.2.

5.1 Conclusions

Object recognition is still a central goal in the computer vision community. In the last decade there has been a huge progress mainly due to the introduction of powerful image descriptors and robust machine learning techniques. Nonetheless, the level of achievement of the developed systems is still far from the human performance that achieves the recognition task in a fast and effortlessly manner. Following this idea, in this work we pursuit to research by getting the inspiration from the human visual system and more concretely, from the ventral stream pathway that is the responsible of the recognition task. Therefore, the methodological assumptions we are adopting for this research is twofold:

- Ventral stream pathway presents a hierarchical architecture that is being assumed as a central property to achieve the abovementioned levels of performance.
- Hierarchical architectures allow working with the abstractions provided by the intermediate levels, providing an interesting methodology towards the definition of visual codes.

As a starting point for the proposed research, in this work we have studied already defined low-level vision models of the first stages of the visual system. To this end, we have reviewed previous works by analysing three low-level vision models (ID,HMAX and MP) and we have unified in a three level frame: L1 (linear filtering), L2 (first non-linearity), L3 (second non-linearity); which matches with the NLN schemes (Normalized Linear Non-linear). With the goal of studying these low-level models, we have decided to work as follows:

- We have selected the ID model as the focus of our study. It uses multi-resolution gabor-wavelet decomposition for L1, a centre-surround enhancement for L2 and a divisive normalisation for L3; by evaluating different alternatives for each stage. Choice is due to its good performance in predicting colour induction and saliency estimation with a unique generic model.
- We have selected saliency prediction as the basic problem to perform our analysis, decision has been based on the availability of large datasets and the existence of an evaluation framework.

From our study we can summarize the following conclusions:

- We have analysed the usual performance evaluation measurements used in saliency prediction and we have identified some negative effects of the KL measure.
- We have proposed a new evaluation measurement, WARP, that tries to overcome the problems of KL that give more credit to more discriminative saliency maps.
- We have studied L1 level by replacing multi-resolution Gabor-wavelet by a DOG filter family; it shows an improvement in the ability of capturing a larger number of features and a more accurate detection.
- We have studied L2 level by introducing shaped windows for the centre-surround mechanism, again this shaped approach provides with a more accurate response and opens the possibility to be adapted to detect more complex features, such as, end-points, bars and blobs, versus the oriented edges provided by the wavelet used before.
- We have studied L3 level by exploring the effect of sigmoid functions instead of the rational quadratic used previously, we have proved that an adequately fitted function for the task can clearly improve the performance.
- The performed experiments appear to diminish the effects of L4 as a weighting stage (ECSF), improvements provided by the refinements studied in the previous levels are reducing the impact of this weighting stage.

Conclusions about the exploration of deriving representations that can drive to visual codes for recognition:

- We have derived a new representation from the model outputs that provides with a visual descriptor oriented to a higher-level task such as face or generic object recognition.
- We have experimented with this new representation as the input for a trainable layer that could fit a weighting function for a higher-level task, but the experiments are already too preliminary to give specific conclusions.

5.2 Further work

Our study leads us to face new issues that could be analysed in further works.
At L1 level, the set of filters used (DOOG) depends basically on two parameters: the sizes and the orientations. An accurate study can give us the best settings to achieve the best results and to put the basis for improving it with specific centre-surround in the next stage.
At L2 level, the centre-surround regions follow a constraint size-relation between them. A multi-scale can improve the method, varying the sizes rate between centre and surround and also, the space-scale construction. Our centre-surrounds regions follow circles and ellipses. An extension to different shapes such as points, bars can be useful to detect more complex responses.
At L3 level, we have observed that the effect of the normalisation function can have important effects, therefore the shape or the parameters of this function requires to be studied depending on the task is being faced.
At L4 level, our hypothesis still remains that ECSF could correspond to a responses task-adaptation step related to the L3 parameters. For saliency estimation, its effect has been reduced. Nevertheless, extending previous layers to other selectivity-responses this function can improved the result and could be learned using common machine learning techniques.
At a global level, although the method shows good results, it does not implement all the feature sensitivity that is assumed to be performed in V1. The model is, then, extensible to complete V1 responses and adding colour, 3D shape, motion,...
At a higher level, the proposed model should be extended beyond V1, facing higher-level visual tasks. After this work we hypothesize that some of the mechanisms explored in this early levels could also be used or improved to be applied in further levels.

Bibliography

[1] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[2] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. Rodriguez-Sanchez, and L. Wiskott, "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1847–1871, 2013.

[3] X. Otazu, M. Vanrell, and C. A. Parraga, "Mutiresolution wavelet framework models brigthness induction effects," *Vision Research*, vol. 48, pp. 733–751, Feb 2008.

[4] X. Otazu, C. A. Parraga, and M. Vanrell, "Toward a unified chromatic induction model," *Journal of Vision*, vol. 10(12), no. 6, 2010.

[5] N. Murray, M. Vanrell, X. Otazu, and C. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 433–440.

[6] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Low-level spatio-chromatic grouping for saliency estimation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, June 2013.

[7] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, vol. 73, pp. 415–34, 2012 Feb 9 2012.

[8] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 411–426, 2007.

[9] T. Serre and T. Poggio, "A neuromorphic approach to computer vision," *Communications of the ACM*, vol. 53, no. 10, pp. 54–61, Oct. 2010.

[10] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 8, pp. 333 – 341, 2007.

[11] S. Fidler, M. Boben, and A. Leonardis, "Similarity-based cross-layered hierarchical representation for object categorization." in *IEEE Computer Vision and Pattern Recognition*, 2008.

[12] N. Pinto, D. Doukhan, J. J. DiCarlo, and D. D. Cox, "A high-throughput screening approach to discovering good forms of biologically inspired visual representation." *PLoS Computational Biology*, vol. 5, no. 11, 2009.

[13] N. Pinto, D. D. Cox, and J. J. Dicarlo, "Why is real-world visual object recognition hard," *PLoS Computational Biology*, 2008.

[14] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *Journal of the Optical Society of America A*, vol. 7, pp. 923–932, 1990.

[15] J. Zhang, Y. Barhomi, and T. Serre, "A new biologically inspired color image descriptor," in *Proceedings of the 12th European conference on Computer Vision - Volume Part V*, ser. ECCV'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 312–324.

[16] D. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neuroscience*, vol. 9, pp. 181–197, 8 1992.

[17] K. Mullen, "The contrast sensitivity of human color-vision to red green and blue yellow chromatic gratings," *Journal of Physiology*, pp. 381–400, 1985.

[18] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," in *AI MEMO*, 2005.

[19] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999.

[20] I. M. Finn and D. Ferster, "Computational diversity in complex cells of cat primary visual cortex," *The Journal of Neuroscience*, vol. 27, no. 36, pp. 9638–9648, 2007.

[21] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, pp. 91–97, 1981.

[22] R. Young, "The gaussian derivative model for machine and biological image processing," *General Motors Research Laboratories*, p. 5128, 1985.

[23] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, pp. 32–32, 2008.

[24] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[25] J. K. Tsotsos and N. D. B. Bruce, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds., MIT Press, 2006.