Modeling Visual Attention

Anna Lázár

A thesis submitted for the degree of
Doctor of Philosophy

Scientific adviser:
Tamás Roska, D.Sc.
ordinary member of
the Hungarian Academy of Sciences

Supervisor:
Zoltán Vidnyánszky, D.Sc.

Faculty of Information Technology
Pázmány Péter Catholic University

Budapest, 2008

Abstract

Visual attention is the ability which allows us to direct our gaze *rapidly* towards *objects of interest* in the visual environment. In this definition *"rapidly"* means *in real time* (it is enough to think of its' evolutionary importance: detecting predators, preys, etc.), whereas the problem concerning what are the *"objects of interest"* in a given moment, is extremely complex: it depends on the systems' actual activity, inner state and different outer conditions as well. In a nutshell, visual attention is a complex and difficult task, which is being performed very effectively by living creatures, whereas it is extremely haltingly imitatable for artificial systems, demanding enormous processing capacity. Mammalian attentional system consists of two different, but closely related parallel working mechanisms: a reflex-like, involuntary one, called "bottom-up" and a volitional one, called "top-down". In the present dissertation I describe the *design*, *realization*, *adjustment* and *testing* of a bio-inspired (partially "neuromorphic") bottom-up attentional model, applying a CNN-based (Cellular Neural/Nonlinear Network) mammalian multi-channel retina simulator. The included parameters have been optimized based on human gaze direction measurements during viewing complex dynamic natural scenes. Similarly, the model's *accuracy* has been determined by comparing its' *predictions* to *measured* human fixation locations. Overall it has performed very well: the measured locations have been among the first four predicted locations in more than 70% of the cases, for which the accidental chance is less than 20%. Finally, I also report on some related practical applications I have realized. These tasks have raised in the "Bionic Eyeglass Project", which aims to help the everyday life of blind or visually impaired people.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preface

Attempts aiming to understand *human vision* dates back to Antiquity. Euclid (˜300 BC) already wrote about perspectivity, that is, how the three dimensional world can be mapped onto a two dimensional surface. But, in spite of some early results, ancient Greeks had a quite incoherent conception about vision. Aristotle thought, that the eye emits light-beams – similarly to distance-indicators used in our days. This idea had been transmitted by Arabic scientists (primarily by Alhazen) to the western world and influenced the conception about vision until the late 16th century. During this period, the fundamental problem was the nature of the physical connection between the objects and the eyes. Worth noting, that although the basic concepts about vision were completely incorrect, based on the given ideas, Alberti and Brunelleschi were able to work out the theory of perspective pictures early in the 15th century, and that spectacles were in common use by that time, notwithstanding the complete absence of any theory to explain why they worked.

According to an alternative idea, conceived during these centuries, the objects send out copies of themselves, which are captured by the eye. Under this hypothesis, the fundamental questions were targeting the physical nature of these copies, as well as the anatomical background of their incorporation.

Finally, the reassuring solutions to these questions have been given by Kepler in the beginning of the 17th century, who provided a theoretical explanation of the optics of the eye in 1604. Soon after, in 1624 Christopher Scheiner in a direct experimental demonstration showed, that an optical image is indeed formed on the inside rear wall of the eyeball. These two events have actually grounded *modern vision research.*

Regarding the problem of *attention* – although the word itself has a Latin root and had already been used in the Ancient world – Descartes is the first who mentions it in a scientific context in 1649 [11]. He related it to movements of the pineal body acting on the animal spirit:

> "Thus when one wishes to arrest one's attention so as to consider one object for a certain length of time, this volition keeps the gland tilted towards one side during that time."

After him, Hobbes, Malebranche and others formulated some interesting ideas, and for the first time, Leibnitz (1765) and Wolff (1734) introduced the idea of *apperception*, by which they interpreted attention as a 'gateway' to consciousness:

> "In order for the mind to become conscious of perceived objects, and therefore for the act of apperception, attention is required."

In other words, attention implements an information-processing bottleneck that allows only a small part of the incoming sensory information to reach short-term memory and visual awareness [6, 12, 13].

In spite of these rather philosophical or sometimes naive explanations of vision and attention, our aim is to find a restricted, however efficient model for some kind of real-time tasks.

Namely, in our days, the phrase *visual attention* is basically used in the following sense: [6, 14, 15, 16, 17, 18, 19].
It is an ability, which allows the (living or artificial) creature to direct its' gaze rapidly *toward* objects of interest *in the visual environment.*

This definition has two important passages: Firstly, *"rapidly"*, which means *in real time.* This becomes obvious, if one thinks about the evolutionary background of visual attention: for a living creature, it is fundamental to detect possible predators in the moment they appear, signals from mates, possible preys etc., no matter how cluttered the visual scene is. Secondly, *"objects of interest"*, which means the ability of identifying those objects or regions in the visual environment, which contain the actually important information. But the problem concerning what is important in a given moment, is extremely complex. It depends also on the systems' actual activity, inner state and different outer conditions as well. To conclude, visual attention is a complex and difficult task, which is being performed very effectively by living creatures, whereas, as it has turned out, it is extremely uneasily imitatable for artificial systems, demanding enormous processing capacity.

Though, despite of all difficulties, visual attention modeling is a very active research area. From an engineering viewpoint, a system that *attends*, determinates the "region of interest" (ROI), possesses very advantageous features, since it does not have to process all the data that is present in the visual environment, but only a small part of it. Thus, the *quality* of the processing can be increased significantly, whilst the *time* necessary for it can be decreased, by omitting the redundant and/or actually unimportant data.

1.2 Main objectives

Mammalian (and thus partially human as well) visual attentional mechanism is composed of two different, but closely parallel working methods: one of them is a fast, reflex-like process, which directs ones' gaze toward sudden, unexpected stimuli, for example toward a flickering red lamp on the street, while the other is a slower, intentional method, which originates in the high brain areas (primarily in the prefrontal cortex). One – for example – is able to find his key in a crowded drawer with the assistance of this latter process, which is called "top-down"(TD), after its' projection direction: it originates in the "top" (high brain areas) and the signal travels "down", that is, towards the eyes. In contrast, the reflex-like method is called "bottom-up"(BU), which indicates that it originates in "low" areas, namely in the retina, and projects towards the cortical regions, where the stimuli is appercepted.

General neuromorphic top-down attention models in our days do not yet exist, primarily because the underlying mechanisms are fundamentally not yet understood: little is known about awareness, conscious decisions and the structures materializing them. In this top-down mechanism cognitive human aspects might play an important role as well. In contrast, the areas to which the bottom-up method can be basically tied to (primarily the retina, LGN, V1 and some other "low" areas) are much better described and understood [20] [21]. Besides, from an engineering viewpoint, the intentional search (TD process) modifies the processes used in the BU method in well-definable points (for example at the weighting of the channel-based saliency maps, see chapter 2.1 and 3.1, [6]). Furthermore, experiences show, that the BU process can be surprisingly effective in many cases, thus many applications can also be based on it. Probably these are the main reasons why the majority of the attempts aim to realize the bottom-up method.

The primary goal of the work being carried in the present dissertation, is to **design**, **implement**, **effectuate** and finally to **test** a *general, partially neuromorphic bottom-up* attentional model, which on the one hand, stays as close as possible to the biological basics in its' structure, and on the other hand, gives verifiable predictions which can be compared to "real" human fixation locations. Of course, an essential goal has been to bring the accuracy of the model as high as possible, where the "accuracy" has been defined by comparing the models' predictions to measured human fixation locations. Two separate sections (3.2 and 4.1) deal with the placement of the introduced model among other approaches by detailing the differences and novelties compared to other models. Additionally, a fundamental objective has been to **apply** the model – or its' parts – in some real-life problems as well.

1.3 Framework of the dissertation

The basic topic concerning the present dissertation is *modeling visual attention*. Figure 1.1 summarizes the framework of the thesis, and in particular, how the different chapters relate to each other (the chapter numbers are depicted with bold italic letters).

The second chapter – right after this introduction – describes the biological background of visual attention: how it operates in living creatures, what the main methods and principles are and details the involved brain areas. This is followed by the corresponding models known from the literature (chapter 3), whereas chapter 4 describes the model I have realized during my Ph.D. years, emphasizing the novelties and differences compared to the existing ones.

The proposed model initially had included two important, yet unknown parameters, namely (1) the receptive field sizes belonging to the different channels, and (2) their channel weights (detailed in section 4.2). The values of these parameters have been estimated via human gaze direction measurements – the set-up of the corresponding experiment is described in the Appendix B. The next two chapters refer to the methodology I have used for the estimation of the quested parameters: the receptive field sizes and the channel weights, in chapters 5 and 6, respectively. Chapter 7 details the validation results of the model adjusted by the previously estimated values. The verification data relies on the comparison of measured human gaze directions to the predictions of the model.

The next chapter (8th) somewhat stands alone: it is about the already existing practical applications, which are tasks raised in the so called “Bionic Eyeglass Project”. Finally, the last chapter summarizes the new scientific results, itemized in the form of separate points (theses).

The basic definitions and abbreviations are collected together in the Appendix C.

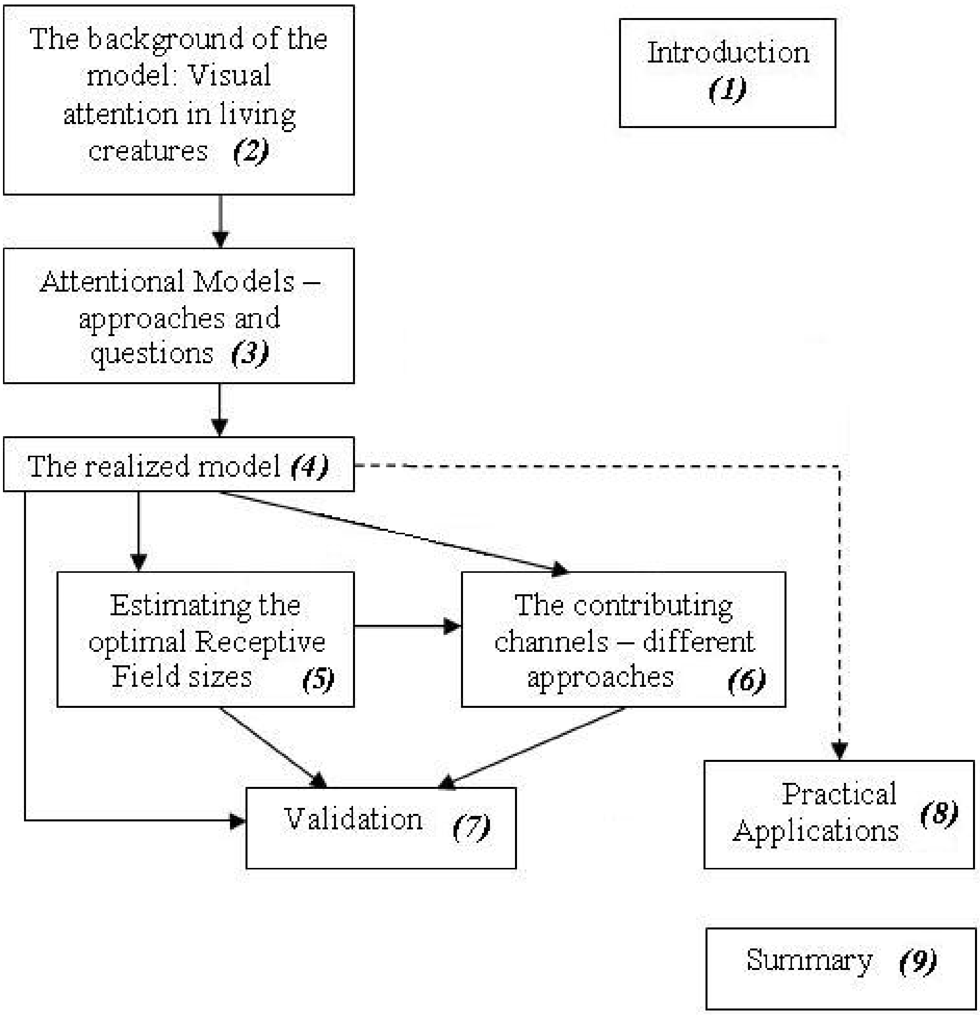
The background of the model: Visual attention in living creatures (2)
Introduction (1)
Attentional Models – approaches and questions (3)
The realized model (4)
Estimating the optimal Receptive Field sizes (5)
The contributing channels – different approaches (6)
Validation (7)
Practical Applications (8)
Summary (9)

Figure 1.1: Framework of the dissertation.

Chapter 2

The Background of the Model: Visual Attention in Living Creatures

Selective visual attention refers to the mechanism by which a creature can direct its' gaze rapidly towards objects of interest in the visual environment. This mechanism allows only a small part of the incoming sensory information to reach the short-term memory and visual awareness [22], permitting, in this sense, the creature to break down the complex problem of scene understanding into a rapid series of computationally less demanding, localized visual analysis tasks [6]. Visual attention is often compared to a rapidly shiftable *spotlight* which scans the visual environment either covertly (when the eyes remain fixed) or overtly (when the direction of the gaze follows the focus of attention).
Of course, vision as a complex task, is not merely attentional, since one can derive a coarse understanding from short appearances of visual scenes being so brief, that they do not leave time for attention to explore the stimuli. Thus, vision appears to rely on an elaborated cooperation of a coarse, massively parallel, full-field pre-attentive analysis system and a more detailed, circumscribed, sequential analysis system [14].

Biologists often distinguish three main functions, which compose visual attention in living creatures [23, 24]:
The first component is the **selection**, which contains those mechanisms that al-

low the creature to *sort out* those sensorial stimuli that agree to the creature's actual purpose.

The second one, **wakefulness**, refers to the mechanisms that *maintain* the continual attentional level.

And finally, the third one is called **control**, which enables the *dynamic* attentional shifts, according to the creature's actual condition, purpose and task.

(Although this chapter is dedicated solely to the biological background, I anticipate that from an engineering viewpoint, only the *selection*, that is the *first group* is important, since the sustaining and the releasing of attentional resources do not appear as a problem in artificial systems. This is the reason why "visual attention" and "selective visual attention" often appear as synonyms in engineering literature dealing with visual attention modeling.)

2.1 Bottom-Up & Top-Down

As mentioned in the Introduction, visual attention basically consists of two mechanisms: a volitional one, which is called "top-down" or "task-dependent", and a stimulus-driven, which is respectively called "bottom up" or "image based" process. Top-down attentional selection is determined by the current goals of the organisms and is mediated by the top-down modulator projections from the front-parietal areas to the visual cortex [25, 26]. For example, searching for a red pen in a crowded drawer will result in a top-down attentional facilitation of the visual cortical neurons, coding the red color and suppressing those which are selective for other colors [27, 28, 29]. It should be noted that in the top-down process there might be a human cognitive aspect involved. Actually, the exact fully neuromorphic model of the bottom-up pathway is not known (not to mention the top-down). On the other hand, bottom-up attentional selection is determined by the physical properties of the visual input. In case of abundant visual input - consisting of many different visual objects - there is a competition between the neural representations of different objects that are simultaneously present in the visual scene. Bottom-up attentional selection refers to the mechanism as a result of which the most salient visual objects of the scene - according to its' physical properties - gain processing advantage and are going to “capture our attention” and evoke an eye movement towards it. For a comparison see table I.

Table 2.1: The main features of the Bottom-Up and of the Top-Down attentional mechanisms; a comparison

Bottom-Up	Top-Down
Image-based	Task-dependent
Originates in the low brain areas (retina) and projects towards the high regions (prefrontal cortex)	Originates in the high brain areas (prefrontal cortex) and projects towards the low regions (retina)
Involuntary (reflex-like)	Voluntary
Takes 25 ˜50 ms	Takes ˜200 ms
It comes before getting aware of the scenery	The process of the visual features can be adjusted voluntarily according to the a task
example: a flickering red point in front of a gray background	example: searching for a key in a crowded drawer

2.2 The Involved Brain Areas

Although the bottom-up(BU) and the top-down(TD) methods work in a strongly parallel way, both processes can be bounded to specific brain areas. The most important brain structures involved in the BU process are primarily the retina (section 2.2.1), the Lateral Geniculate Nucleus (LGN), the Colliculus Superior (CS) and the V1 (section 2.2.2). From that level, the brain areas are more tightly bounded to the TD process. From V1 the processing dissolves into two main pathways: one is the “where pathway”, which comes from the periphery of the retina and goes towards the posterior parietal cortex, and the other is the “what pathway”, which comes from the fovea and directs to the infero-temporal cortex. The former participates in the movement and spatial information processing, while the latter contributes to the object recognition - of course, in a tight cooperation (section 2.2.3).

During the discussion of the visual process, it is important to keep the three main organizing principles in view, which characterize the entire visual processing and structure. These are:

- Topography
- Parallel processing
- Hierarchy

2.2.1 The Retina

The retina is a multi-layered sheet of nerve cells at the back of each eye which converts light into electrical signals, which are transmitted to the brain through the optic nerves and tracts (figure 2.1). Since its' functioning and structure has a fundamental role in the developed model, it is expedient to devote a separate section for the structure and the functioning of the living retina.

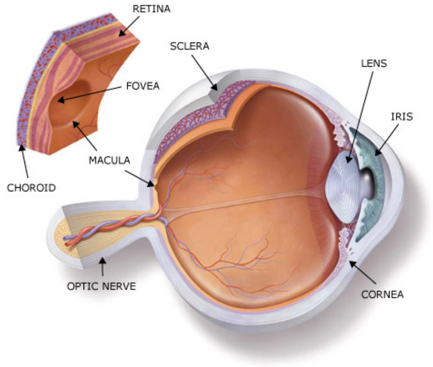

Figure 2.1: The structure of the retina: it is a multi-layered sheet of nerve cells at the back of the eye which converts light into electrical signals that are transmitted to the brain through the optic nerves and tracts. The picture is from [1]

Mammalian visual system perceives the outside world through several different channels. These spatio-temporal channels arise in the retina and persist until the high brain areas - whilst several processing steps occur on them. The question referring to how and where do these unite into a uniform visual perception is still open. Between the photo receptors (which intercept the photons) and the ganglion cells (which axons form the eye's "output", the optic nerve) there are several layers and cell-types which already start to process the information in the retina. The retina has ten histological layers. (figure 2.2) The information flows through the *vertical pathway* composed by the *photoreceptors*, *the bipolar* *cells* and the *ganglion cells*. Among these layers the two synaptic strata lie: the *Outer Plexiform Layer (OPL)* between the photoreceptors and the bipolar cells, and the *Inner Plexiform Layer (IPL)* between the bipolar cells and the ganglion cells, which are ordered stacks of synaptic planes. These strata primarily do not *convey* the information but they *modify* it [20][30].

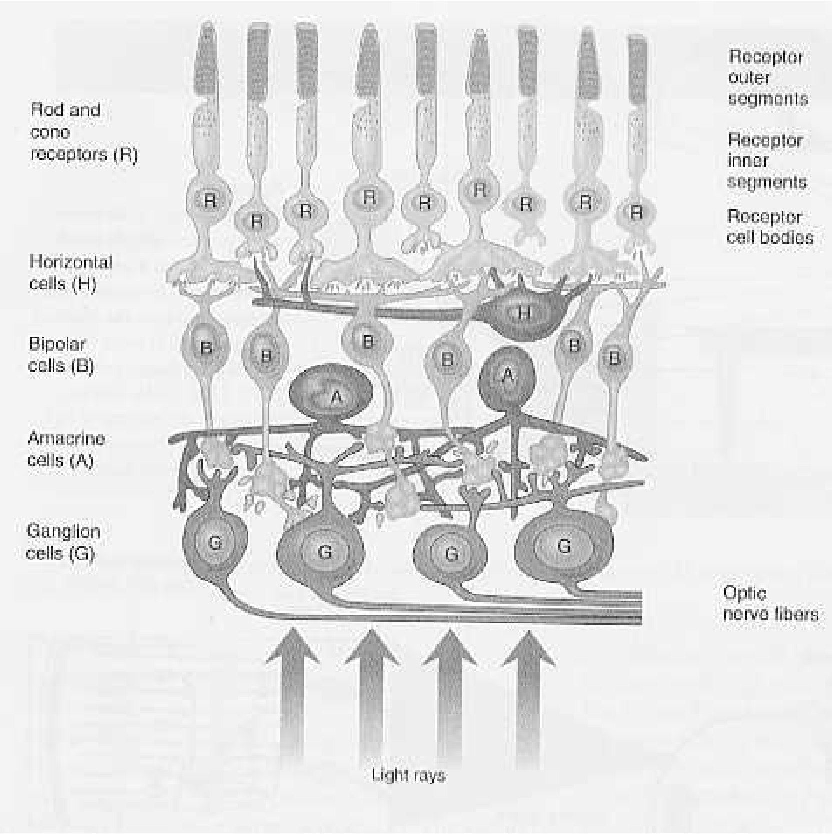

Figure 2.2: Organization of the retina in a schematic vertical view. In reality the retina is packed with cells, and there is hardly any extracellular space. The "vertical pathway" is composed by the photoreceptors (rods and cones), bipolar cells and ganglion cells, among which layers lie the two synaptic strata: the Outer Plexiform Layer between the photoreceptors and the bipolar cells, and the Inner Plexiform Layer between the bipolar cells and the ganglion cells. The picture is from [2].

In the first step light is captured by the photoreceptors: rods and cones. In daylight cones are active, while in dim lighting conditions the rod-system works. Nevertheless, in most of the mammalian retinas rods outnumber cones by around

20-fold, there are usually about 8-10 times more neurons in the retina driven by cones than by rods [30].

The next cell-group of the vertical pathway consists of the bipolar cells, of which a typical mammalian retina has around 9-11 different types of. These cells connect the inner and the outer retina. They are explicitly oriented: the dendrites always go towards the photoreceptors while the axons branch in the inner plexiform layer. The approximately ten bipolar cell types define the decomposition of the visual information: they compose different channels, which join to the different ganglion cell types selectively. Bipolar cells stimulated by cones are usually called "cone-bipolars". All cones release glutamate, but different bipolars react differently to this neurotransmitter: about the half of them has ionotropic glutamate receptors: these get depolarised by glutamate through a cation channel, while the other half has sign-inverting synapse: these get hyperpolarised through metabotropic glutamate receptors (mainly mGluR6).

Since photoreceptors get hyperpolarised by stimulation (practically: light), those bipolar cells that respond to stimulation with hyperpolarisation are *sign conserving*: these are the *OFF-cells*, while those that get depolarised by the photoreceptors hyperpolarisation are *sign-inverting*: these are the *ON-cells*. This distinction, which has evolved in the first retinal synaptic step, remains throughout the whole visual system. Furthermore, both the ON and OFF classes subdivide into further distinct classes according to the response-time: there will be separate channels for high-frequency (transient) and low-frequency (sustained) information. The individual bipolar cell-types branch at different layers of the IPL, where they find different amacrine and ganglion cells as possible synaptic partners.

The output of the retina is formed by the axons of the ganglion cells. In mammals there are approximately a dozen types of them, which can be classified by dendritic arbor, structure, physiology, and branching level. The dendrites of the different ganglion cell-types ramify at distinct strata of the inner plexiform layer, and they embody different representations, features of the visual world [21]. The width of the dendritic arbor is the area that the given cell is able to perceive; this area, that is the patch of the visual field that any single neuron monitors, is called that cell's *receptive field* (RF). On the layer of the ganglion cells appears the central-peripheral organization of the RFs. In this, a circle-shaped central part is surrounded by an antagonistly responding peripheryal part. Homogeneous light covering the whole RF results in no response (figure 2.3).

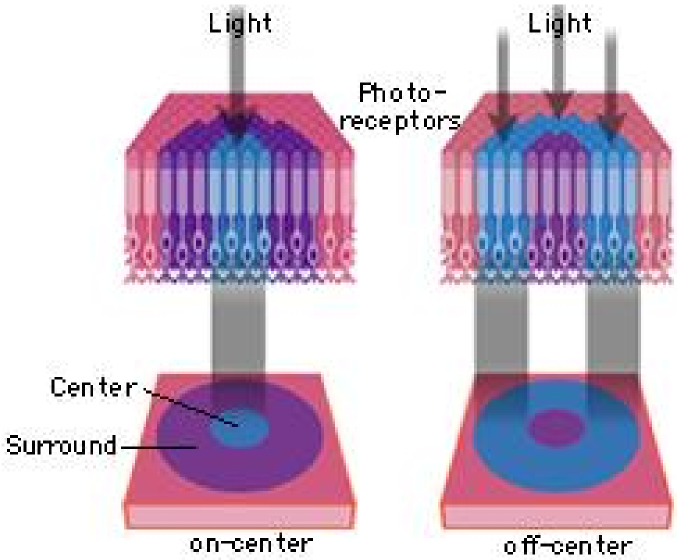

Figure 2.3: The patch of the visual field that any single neuron monitors is called the cell's receptive field. Center-surround receptive fields arise from a pool of photoreceptors. The photoreceptors can either act to excite or to inhibit a downstream cell. In an on-center bipolar cell, light hitting the central photoreceptors will be excitatory and light in the surround will be inhibitory. In an off-center bipolar cell, light in the center will be inhibitory, and light in the surround will be excitatory [3].

2.2.2 Between the Retina and the V1

After the nerves leaving the eye cross in the chiasma opticum, (in humans about the 50% of the nerves changes brain-side here) about 80% of the fibre projects to the thalamic nucleus LGN, while the remaining 20% goes to different midbrain structures from which the Superior Colliculus (SC) is the most important (figure 2.4). These two visual centres are rich in different connections with other areas. The SC is an ancient, more primitive area than the visual cortex. Besides the visual stimuli, this area receives audio input as well. If the audio and the visual stimuli appears at the same time, the generated response is much stronger. The principal role of this structure is supposedly to detect those objects, which are at that moment further from the fixation point, but otherwise could be important for some reason, and to direct the gaze towards them. Nevertheless, the SC is not able to process vision in detail [31, 32].

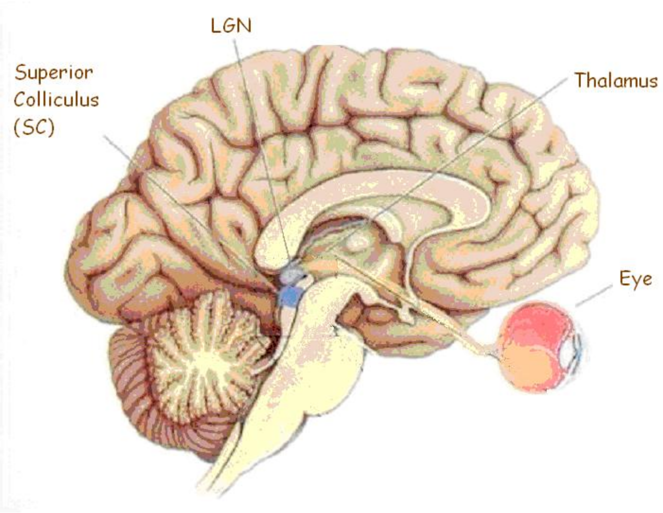

Figure 2.4: Brain areas involved in vision and visual attention between the retina and the Primary Visual Cortex (V1). After the nerves leaving the eye cross in the chiasma opticum, most of the fibers project to the LGN, while the remaining go to different midbrain structures, e.g. to the SC. Axons of the LGN cells project to the V1.

The LGN consist of (in humans six) different layers. All of the layers have a *topographic map* of the visual scene, which means that those points that are adjacent in the outside world are represented by adjacent neurons in the LGN as well. This topographic alignment is typical in the whole sensory system, from the retina up to the high brain areas. In the LGN, the six topographic maps are located in a way that the similar parts of the different maps lie right under each other. Similarly to the retina, circle-shaped receptive fields belong to the LGN cells, organized into oppositely responding central and peripheral parts. Similarly to the topographic alignment, the receptive field organization is a basic principle in the neuromorphic signal processing as well. An important difference between the LGN and the retina is that the LGN's peripheral cells inhibit much stronger than those in the retina, adding up in a more emphasized contrast.

From here, the axons of the LGN cells project to the V1, where more complex processing begins. Important to note, that the direction of the information-flow is not unidirectional, because the LGN receives surprisingly many signals from the V1, which indicates significant top-down influence [33, 34]. From here, the information process persists until the prefrontal cortex through several steps.

2.2.3 Higher Areas

Figure 2.5 shows the two main pathways starting from V1 ("Visual cortex 1", "Striate cortex" or Brodmann area 17) [35][4][20]. Functions involved in object *recognition* can be binded to the "ventral stream" (the lower pathway on the figure), which is hence also called "what stream". The process taking place here is slower and more detailed than in the 'Dorsal stream'. The **STS** (**S**uperior **T**emporal **S**ulcus), is the location of face recognition [4]. Many textbook include this into the Ventral stream. Other important areas within the ventral stream are the **I**nfero**T**emporal cortex, (**IT**), which codes object features. Individual IT neurons show preference for a particular pattern. Neurons that code for similar objects or object features are organized into columns. **P**arahippocampal **P**lace **A**rea (**PPA**) responds to places, **E**xtrastriate **B**ody **A**rea (**EBA**) responds to bodies, **L**ateral **O**ccipital **C**omplex (**LOC**) responds to objects. The **F**usiform **F**ace **A**rea (**FFA**), together with the STS, responds to faces as well.

Spatial perception takes place in the Dorsal (or "where") stream. The process here is faster, but as the same time, more rough. One of the most prominent areas specialized for analyzing visual motion, is the **MT** (Motion Area, V5). Without this region the automatic perception of motion is lost. Instead, the visual motion becomes a series of stills, simple judgments of an object's speed and direction become difficult.

Other important areas within the Dorsal pathway are the **A**nterior **I**ntra**P**arietal cortex (**AIP**), which coordinates grasping, **L**ateral **I**ntra**P**arietal cortex (**LIP**), for eye movements, and the **P**arietal **R**each **R**egion (**PRR**), which assists in the action of reaching an object.

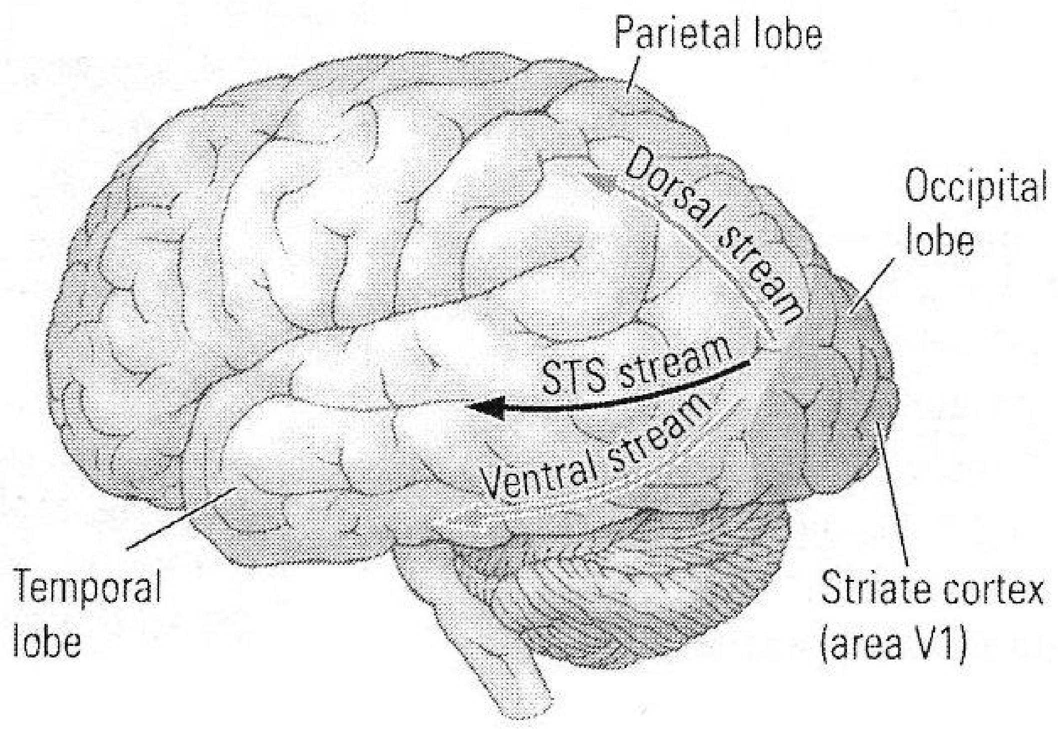

Figure 2.5: The two main pathways contributing in object recognition and localization: the dorsal pathway, that is the "where stream", and the ventral pathway, which is also called "what stream". (The picture is from [4])

Chapter 3

Attentional Models - Approaches and Questions

The first models of visual attention have been developed in the 1980s, after Treisman and Gelade have proposed their feature integration theory [36], wherein they have suggested that only the basic visual dimensions (such as color and orientation), the so-called 'low-level visual features' are processed throughout the visual field in a parallel way. Afterwards, it is the visual attention that binds together the low-level features belonging to the same object into coherent object representation. The later, attention-based process takes place in a serial way; attention is allocated to one or at most a few objects at a time.

A detailed BU, stimulus-driven visual attentional model has been proposed by Koch and Ullmann in 1985 [5]. In this model feature-specific 'saliency maps' have been calculated for the different visual features (color, orientation, etc.). 'Saliency maps' are scalar, two-dimensional topographic maps, representing feature contrasts rather than a given feature's absolute value, at each location of the visual field. As a next step, feature-specific saliency maps have been integrated into a so-called 'master' or 'final' saliency map. In the master map the saliency representation was already feature independent. Lastly, due to a 'winner-take-all' mechanism, the most salient part of the master map (which has the highest salience value) gains processing advantage and captures attention, while other salient parts of the map are suppressed.

Osberger and Rohaly have identified some factors on complex scenes, which have strong influence on visual attention [37]. Based on these, they have created a

model that is able to make predictions for human gaze directions. Most of these features were driving the BU process (motion, contrast, etc.), some of these were related with the TD process (people, context), while some were in ‘between’ (shape, foreground/background distinguishment). They also highlighted the difficulty of the weighting of these features.
In the last two decades, several models of visual attentional selection have been developed [38, 39, 40], most of them sharing the main components of the original Koch and Ullmann model [41][38]. There are some important characteristics of these models:
(1) the choice of the low-level visual features is heuristic and it primarily depends on the purpose of the given model [6];
(2) the weighting of the individual feature-specific saliency maps during integration into a master map is based on TD approximations, mixing biological findings with heuristic methods to achieve higher efficiency;
(3) with a few recent exceptions[42][43], the models have been tested on static, non-dynamic visual input.

In comparison, I have been primarily focusing on the elaboration of the BU process, taking carefully into account all the features that might have any effect on the bottom-up process. This is being achieved by including all the retina channels – instead of the heuristic low level visual feature extraction – both those whose function is well understood and also those whose function is not sufficiently illuminated up to present. I have managed to give a satisfactory approximation on the weightings of all these features as well. The model has been adjusted and validated on moving input, via human gaze direction measurements.

3.1 The skeleton of a general neuromorphic visual attentional model

Most of the models that work out BU mechanism use more or less the same principles. First, that a point’s final saliency value is composed of several conspicuous-values [44] – each of these belong to different low level visual features (– these are the “feature dependent saliency values”) [6, 38, 45]. Second, that a location’s saliency-value basically depends on the surrounding context, that is, it is not equal with the 'loudness' in an absolute value, but it is proportional with the *contrast* it composes with its' near surrounding [46, 47]. Third, the final saliency map is being aggregated from the feature dependent saliency-values, with different weights. The weighting vitally depends on top-down modulation [28, 48, 49] and can be influenced through training as well [45, 50, 51, 52]. Fourth, scene understanding and object recognition tightly interplay in gaze-direction [53, 54, 55, 56, 57, 58, 59].

Figure 3.1 shows the main steps, which are the followings:

- Dissolve the incoming picture according to low level visual features: colors, intensity (on, off, etc), orientations ($0°$, $45°$, $90°$, etc), motion, junctions, etc. Usually the certain models employ a few of these features, chosen according to their relevance in the given approach or task.

- Create the saliency maps to each channel. There are several strategies, the relevant precept is to measure the contrast between a point and its' surrounding. This is often some kind of 'competition' among near points, which can employ long-range connections as well.

- Feature combination. Unify the feature-based saliency maps into one final one, which is thus already feature independent - ("feature independent" in the sense, that it depends on the whole collection of features). The weighting of the different channels are usually not equal, it is generally under some kind of top-down modulation.

- Determine the most salient point (find the location that has the highest saliency value). This is a winner-take-all mechanism, which means that the whole process was for locating this single point, which will be the attended location.

- Particularly for still images: creating a mechanism called "inhibition of return" which is for preventing attention to rut into one point [60]. This inhibits the system to return to the attended locations for a while, thus attention can move to the next most salient location, then to the third one, etc. This process can also differ in several items in the certain models.

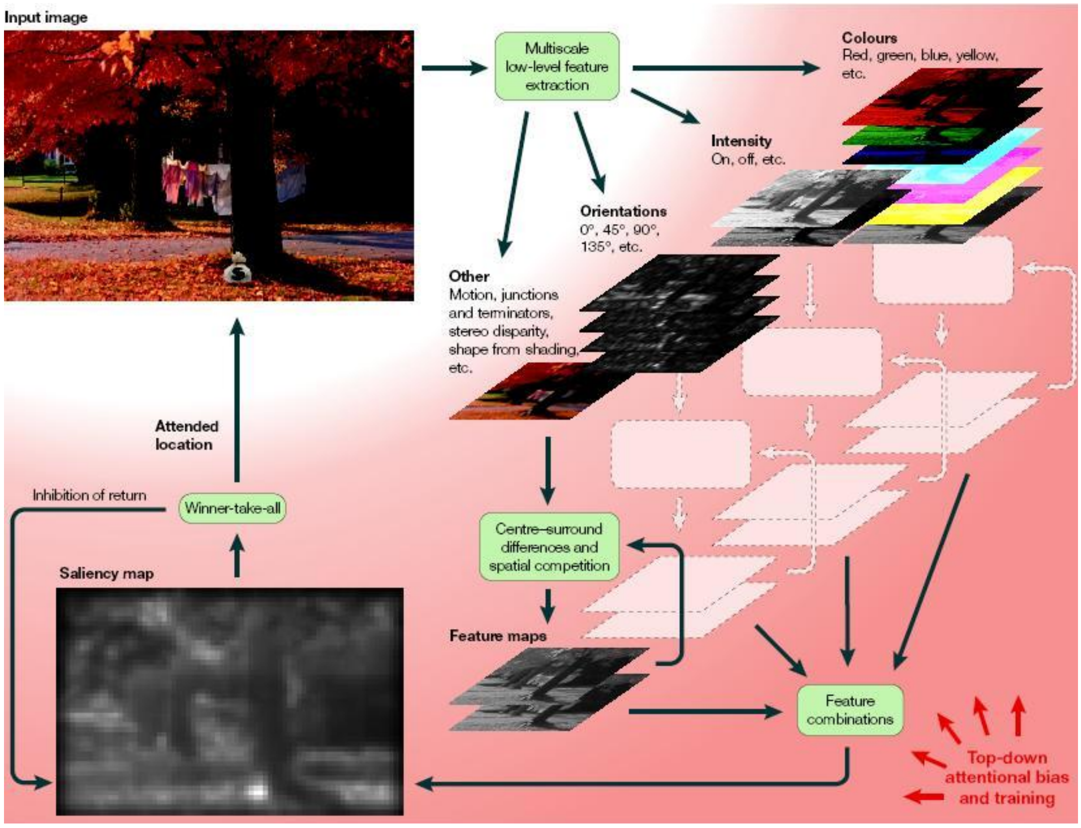
Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre–surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training

Figure 3.1: The skeleton of a general bottom-up attentional method, originally proposed by Koch and Ullman [5]. The picture is from [6]. The input picture (left hand-side, top) is getting dissolved according to the low level visual features (right hand-side, top). Each of these channels create a topographic feature map, which codes the center-surround differences, according to the given features (right hand side, bottom). Then, these maps are aggregated into a 'final saliency map'(left hand side, bottom), each with a certain, mostly top-down dependent weight (right hand side, bottom). The most salient point of the final saliency map attracts the attention, which is suppressed after a while by the method called "inhibition of return" (left hand side, middle).

3.2 'Low-level visual features' contra retina channels

As described in [21], it has been recently discovered that a mammalian retina has ten parallel channels[1], and also the neuromorphic structure of these channels has been found. These channels give qualitatively different answers to the same input. The main differences lie both in spatial and temporal properties. For more biological details I refer to [21, 20, 30] and section 2.2.1.

By these measurements, that have been made on rabbit retina, a first and rough approximation has reached completion and not the detailed circuitry. At the same time, using these findings, this is the first time that we have the possibility to consider this multi-channel pre-processing step. In addition, this step could help solving the biggest difficulty that image processing algorithms nowadays face, namely, that the intensity or color values of the same object can vary in a very large scale according to the scaling and actual lighting conditions, that is, from accidental conditions. Thus we have found the adaptation of this multichannel pre-processing step in attention modeling fundamental. As far as we know, this has been the first attempt to use this bio-inspired channel decomposition in attention-modeling. Moreover, using functional spatial temporal models instead of input-output models in the first part enhances the success of model identification.

This section deals with the function of the different retina channels, meanwhile some details about the used multi-channel retina simulator can be found in Appendix A. As we will see, up to present only half of the channels' functions are known, in the sense that the aim of the process of the remaining five channels could not yet be formulated explicitly. Since these channels do form saliency maps as well – and thus they do take part in the formation of the final saliency map – neglecting them significantly modifies the final results. This has an important corollary compared to models using heuristic low level feature extraction: In

[1] According to the latest researches, certain cells in the retina respond to motion direction-dependently, that is, in certain living creatures, another channel could exist, which filters motion in a direction selective manner [61].

my model, similarly to living systems, the saliency maps that are based on those retina channels having non-explicitly described functions, also take part in the allocation of the fixation location.

Comparing some functional characteristics of three retina channels

During the first year of my Ph.D. studies, on a retina model based on the same principles than described above, we were investigating and comparing the functioning of three retina channels. The name of the simulator of which the measurements had been made on is "RefineC". The results have been reported in [7], the measurements (that are described in the present subsection) had been made together with Robert Wagner, David Balya and Tamas Roska. The investigated channels were:

- **L**ocal **E**dge **D**etector, (LED)
- "Sluggish", which is mostly referred to as "Delta" throughout the present dissertation (since there is not yet a commonly accepted nomenclature for the retina channels, small variances can occur among different papers).
- Transient channel

The basic experiment aiming to clear up the function of the different channels had been a flashing square: a square that appears for one second in a smooth gray background, and then disappears. We had two main reasons to choose this stimulus: firstly, we had the results of biological measurements for the same input [21], with which we could control our results. This gave trustiness for the simulations. And secondly, this stimulus is simple, but at the same time shows the main characteristics.

The results can be seen on figure 3.2: the first row shows the original input, the second row depicts the answer of the Transient channel, the third one is the Local Edge Detector, and the last row is the response of the "Sluggish" (or

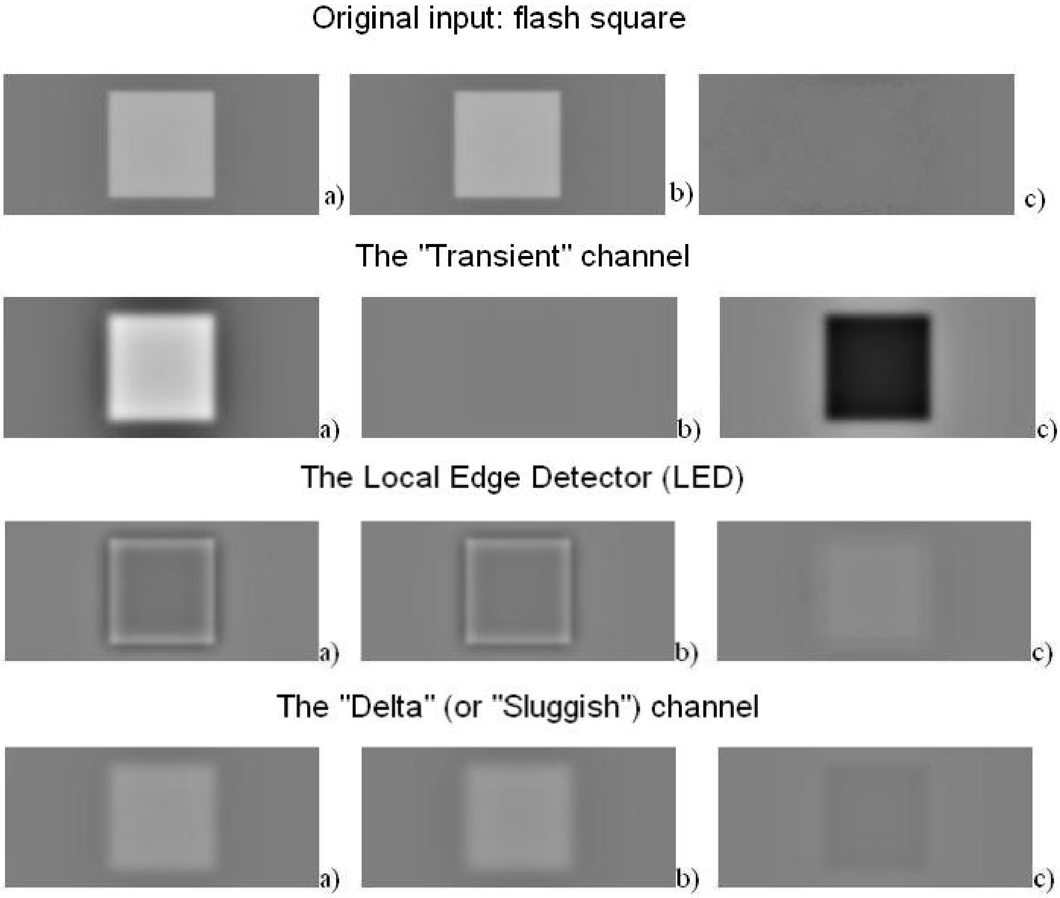
Original input: flash square
a)
b)
c)
The "Transient" channel
a)
b)
c)
The Local Edge Detector (LED)
a)
b)
c)
The "Delta" (or "Sluggish") channel
a)
b)
c)

Figure 3.2: A basic experiment for investigating the retina channel functions: a flash square appearing for one second. The top-most row depicts snapshots from the original input, the other three rows belong to different channels: Transient, LED and Delta (or Sluggish), respectively. All the (a) pictures correspond to the moment when the stimulus pops up. The (b) pictures show the sustained responses, meanwhile the (c) figures correspond to the moment when the stimulus disappears. The Transient channel reacts vividly to all kind of *changes*, LED emphasizes edges, and the sluggish channel "fills out" the objects. [7]

"Delta") channel. Pictures under each other indicates the same moment after the appearance of the stimulus: *a)* in every row is the moment of the pop up, *b)* is at half second, and *c)* is the moment of the disappearance in all the four rows.

The transient channel filters out *changes* and stays silent in the motionless regions (see also figure 3.3, second image in the first row). As we will see, this channel strongly interplays in directing visual attention. The LED channel (third

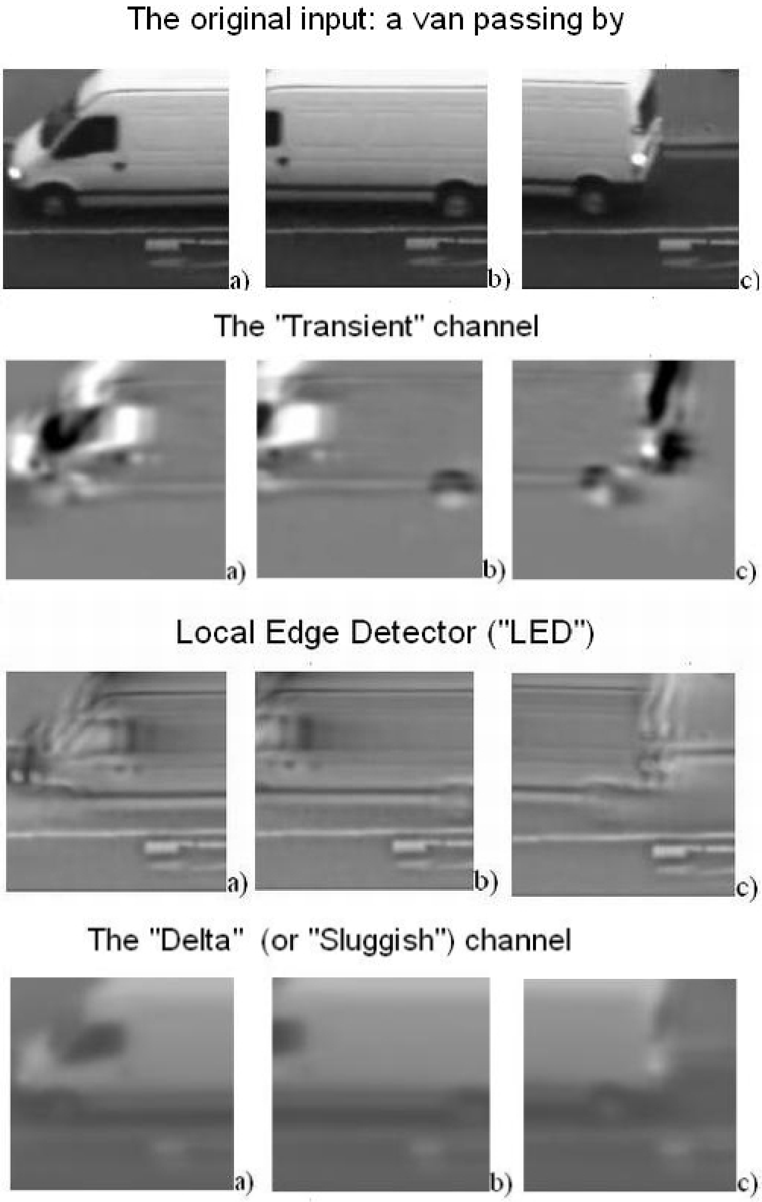
The original input: a van passing by
a)
b)
c)
The "Transient" channel
a)
b)
c)
Local Edge Detector ("LED")
a)
b)
c)
The "Delta" (or "Sluggish") channel
a)
b)
c)

Figure 3.3: The original input is a van passing by. (a) the moment it enters the visual field, (b) it fills out, and (c) leaves the scene. First row: original input, second row: the answer of the Transient channel, third row: LED, and the last row is the response of the Sluggish channel. In contrast with the LED, the Transient channel eliminates the hard shoulder and the strips on the roadway, because they are motionless. The Sluggish channel keeps the shade of the van. [7]

row) also gives vivid answer on motion, but in contrast with the Transient channel, the LED gives response to *motionless* edges as well ("b" pictures) meanwhile eliminating the inner part of the square. This phenomena can be seen even better on figure 3.3, which depicts a van passing by the visual scene. The structure of this figure is similar to 3.2: the first row includes snapshots of the original input, the second row is the response of the Transient channel, beneath the LED, and the last row shows the outcome of the Sluggish channel. As it nicely appears, the Transient channel eliminates the hard shoulder and the strips on the roadway, because they are motionless. In contrast, they do appear on the response of the LED channel. Both the LED and the Transient eliminate the inner part of the objects, if they do not have any special pattern inside (that is: moving edges). This can be observed in both cases: the square and the van.

The reason why we *do* see the shades of the objects and not only their contour, is because of the existence of channels similar to Delta: it seems, that these channels "fill out" the objects around us. (Bottom-most rows in figures 3.2 and 3.3)

Another well-known phenomenon is that we are able to see *coherent motion*, even if it is made of "noise" in the sense that every *snapshot* of the vision is pure noise alone, but as a *flow*, it consolidates into a cohesive, perceptible moving formation. An example for this can be seen on figure 3.4: the first picture (a) is a snapshot of the input video flow: this frame is basically *noise* in itself. Picture (b) depicts the outcome of the Transient channel for the same frame of the video flow: as we can see, the motionless part disappears, while the moving figure, a horse nicely traces out. As we have already seen, the LED channel (c) is also sensitive on motion, but in a different way: the pattern of the background has not vanished completely, because it contains edges, but at the same time, the *inner structure* of the moving object (the horse) has not been kept as intact as in the case of the Transient channel.
This example represents the significance of the *temporal* processing that most channels in living creatures perform (in our model, seven out of the ten: the ones

operating on the discussed retina simulator). Temporal processing enables to retrieve relevant information that actual data solely does not even contain. These channels sometimes are referred to as channels with “memory”.

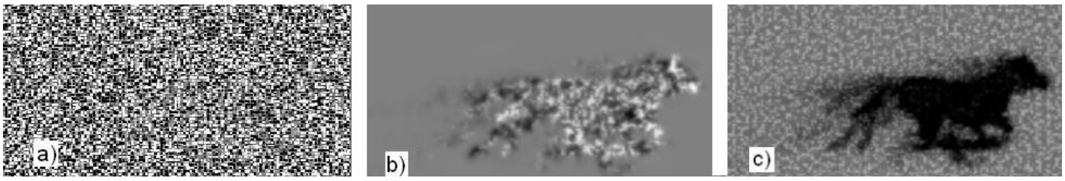

Figure 3.4: Channels with “memory”. The first picture (a) is a snapshot of the input video flow: this frame is basically *noise* in itself. Picture (b) depicts the response of the Transient channel for the same frame: the motionless part disappears, while the moving figure, a horse traces out. The LED channel (c) is also sensitive to motion, but in a different way: the pattern of the background has not vanished completely, because it contains edges, and the *inner structure* of the moving object has not been kept as intact as in the case of the Transient channel.

Chapter 4

The realized model

4.1 The skeleton – summarizing the main steps

This section outlines the main steps of the bottom-up visual attentional model I have realized. In the same time, I emphasize the occurrent differences at each step, compared to the models described in section 3.1. Figure 4.1 illustrates the text.

- The first step is to dissolve the incoming stimuli according to low level visual features. As detailed previously, I have built the model on a multi-channel retina simulator. (see section 3.2 and Appendix A). In half of the channels we can denominate their function, whereas the other five channel's function is still unknown - at least we can not phrase it. Therefore these channels (Polar-, Alpha-, Beta-, Delta- and the Bistratified channels, see Fig. A.2) have never appeared in heuristic artificial models.

- Create the saliency maps referring to each channel. There are several strategies, of which I have used different sized, circle-shaped receptive fields (RF), on and off (section 6.1). Since different receptive field sizes generate different saliency maps on the same input, and also, the extent of these RFs are unknown for the certain channels, these fundamental values had to be estimated somehow. On figure 4.1 red question-marks indicate the unknown parameters.

- Feature combination: unify the channel-based saliency maps into one final one, which is thus already feature independent. In other words, the final (or "master") saliency map is a combination of the channel-based maps, thus it does not depend on only one or a few features, but on all of them. The weights of the different channels are not equal, but the exact ratio is unknown (red question mark on image 4.1). I have investigated different approaches to estimate these weights: "constant" and "continually updated":
  - Continually updated channels weighting strategies: for every frame I have approximated the average and the maximal saliency values appearing on the individual channels, and I supposed that only the first few most salient channels participate in the generation of the master map with weightings that are proportional to their approximated saliency values. The effect of the other channels – on this specific stimulus – is negligible. The exact method how the saliency values had been defined is described in section 7.
  - On the contrary, by constant channel weighting strategies, I have presumed that the different channels participate in the formation of the master map with a pre-defined, invariant ratio.
- Once the master saliency map is ready, the next step is to determine its' most salient point (find the location that has the highest saliency value). This is a winner-take-all mechanism, which means that the whole process aims at locating this single point, which will be the attendant location.
- The "inhibition of return" mechanism, which aims at preventing attention to get stuck into a point, comes to fruition spontaneously, because I am working with moving input, thus the saliency maps change permanently.

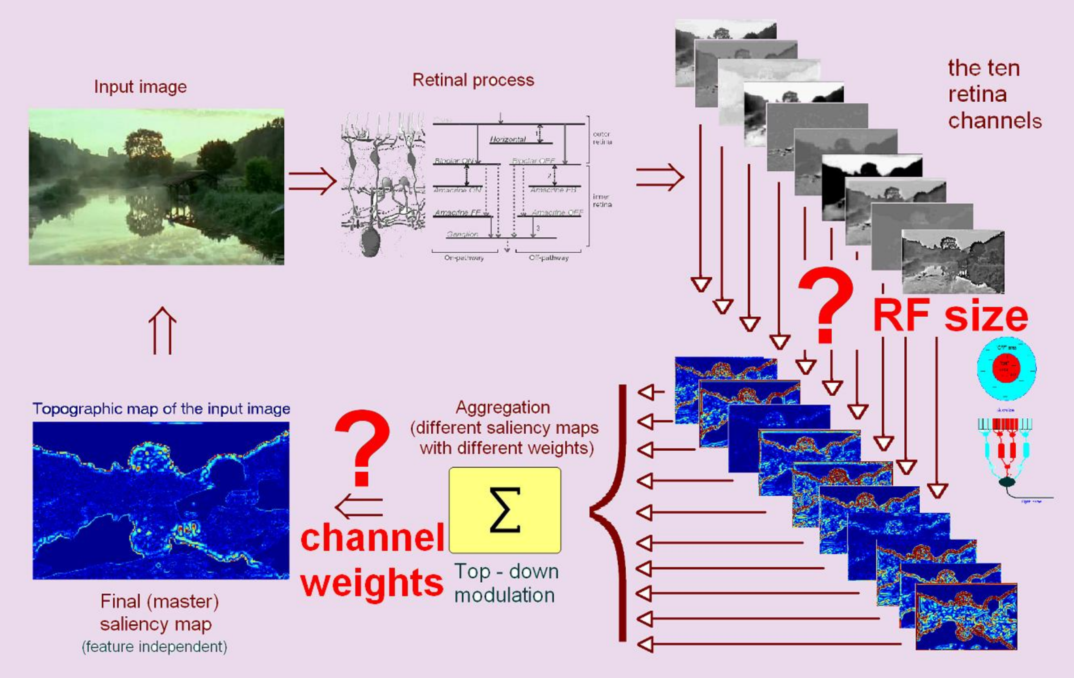
Input image
Retinal process
the ten retina channels
?
RF size
Topographic map of the input image
?
channel weights
Aggregation (different saliency maps with different weights)
Σ
Top - down modulation
Final (master) saliency map (feature independent)

Figure 4.1: The diagram of the bottom-up mechanism I have realized. In the first step the input image (top of the picture, left hand-side) is decomposed into ten different retina channels (– topographical maps in the different brain areas: the higher activity a neuron shows, the darker/lighter color on the monitor appears. This is because the ON channels' and the OFF channels' responses are visualized on the same pictures.) In living beings this is a pre-attentive feature extraction mechanism which operates over the entire visual scene in a highly parallel way. Onces the input vision is decomposed, each retina-channel creates its' own saliency map. For defining the individual point's saliency value, I have used different sized, circle-shaped receptive fields (RF), on and off. The next step is the aggregation. The final (or master) saliency map is practically a weighted sum of the feature-based saliency maps. The weighting of these feature-dependent maps are under top-down modulation, if it is present. (Bottom of the picture.) Then the winner-take-all mechanism chooses the final saliency map's most salient point: this point wins the attention, the others are suppressed. The corresponding picture-portion 'appears in the fovea', this is the small part of the visual scene that is processed in detail and the rest is being processed only roughly.

4.2 Unknown parameters: receptive field sizes and channel weights

Using the model described above, two fundamental questions arise:

1. What should be the EXACT SIZE OF THE RECEPTIVE FIELDS we apply on the different retina channels? Since different RF sizes eventuate in different saliency matrices, this parameter essentially influences the results. In living creatures, it is likely enough that the different channels have a *range* of receptive field sizes. In other words, probably not one single RF size corresponds to the individual channels, but a *distribution* of them. At the same time, probably each channel has a "preferable" size, which dominates. My aim has been to find these preferable sizes for the individual channels, since defining the saliency maps according the a whole range of RF sizes would need unmanageable amount of calculation during the operation of the model.

2. What kind of WEIGHTING STRATEGY should be followed during the creation of the master saliency map? (How do living creatures determine their corresponding topographical maps during bottom-up attentional conditions?) Is it always the same (under BU conditions) or it differs according to the actual stimuli?

Referring to these questions, a fundamental difficulty had been, that *directly* we can *not* measure any of the parameters appearing during the process (for example a saliency map, a channel weight, etc.), but only the *fixation locations* provided by the experimental subjectives. In other words, the process detailed on figure 4.1 is similar to a "black box" experiment in the sense that we only know the input and the output, but we can not measure anything in between. (It is quite difficult to design an experiment, a "stimulus", which affects only one of the channels – it is enough to mention, that if the stimulus is for example dynamic, then it immediately affects the seven spatio-temporal channels and the Intensity one as well.) Accordingly, one can only infer, deduce these directly immeasurable values using different *assumptions*, indirect methods. Furthermore – since for estimating either of the missing parameters, one needed the other – they had to be estimated in an *"interconnected", "coupled"* way. (In this sense, the problem was a bit similar to a Diophantine equation, except that here the criterion was not valid, that the solutions have to be whole numbers.) The assumptions I have used because of the above difficulties will be detailed in chapter 6. After the selection process of the channels triggering the individual saccades – that is, those channels satisfying the criterion appearing on the following diagram (top of the picture) – the corresponding data have been stored and processed (the channel's saliency values in the measured fixation location; one value for each RF size).

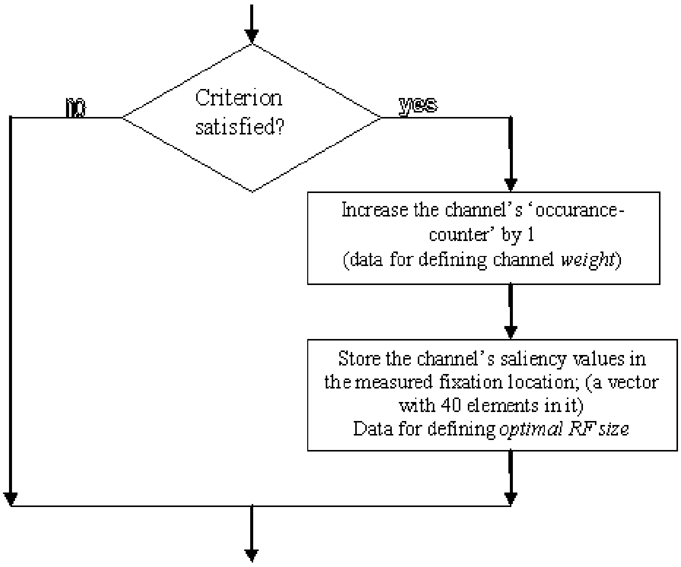
Criterion satisfied?
no
yes
Increase the channel's 'occurance-counter' by 1
(data for defining channel *weight*)
Store the channel's saliency values in the measured fixation location; (a vector with 40 elements in it)
Data for defining *optimal RF size*

Figure 4.2:

Since only those channels' data are taken into account on a given stimulus, which satisfy the criterion of being saccade-triggering – being moreover the criterion distinguishing the different approaches from each other in section 6 – the channels do influence each other via this competition for being treated as saccade-triggering ones. After this step, the different channels 'calculate' their optimal RF sizes independently from each other. In contrast, the channel weights reflect exactly the result of this competition.

(Theoretically, it is possible, that a channel will never satisfy the criterion for being a saccade-triggering one, and thus it will not have, neither a weight, nor an optimal RF size, because there wouldn't be any data to calculate them from, but since this case has never happened (in any of the approaches), I do not deal with it.)

Chapter 5

Estimating the Optimal Receptive Field sizes

One of the primal aims has been to determine the receptive field distribution – and hereby the “optimal RF size” – for all the ten channels. For estimating these values, I have prepared functions representing the *average saliency values* at the attendant locations *in the function of the RF sizes*, for all the ten channels (like on figure 5.4). These diagrams show how “effective” are the individual RF sizes on the different channels at the measured fixation locations. (The way I use the notion of “efficiency” is explained bellow, in the comments.)
During the measurements I have exploited the finding that under pure bottom-up (or “image-based”) conditions, it is the different locations’ *saliency values* that determine gaze direction ([5, 62]).

Some comments:

- During the evaluation, I have used only the *first* measured fixation locations after each of the saccades, because these are the most BU-modulated ones ([43]); afterwards – due to scene understanding and analysis – the TD effect increases.

- Only those saccades had been taken into account that expand at least 1° viewing angle.

- I have investigated 40 different receptive field sizes, spreading from 0.5° up to approximately 26°, expressed in terms of the viewing angle (see figure 5.3). This covers all the RF sizes that are present in the visual system.

- The "efficiency" of a receptive field (size) plays an important role in my approach. I use this term in the following sense: An RF is "efficient", if the biggest values of the saliency matrix belonging to it overlap with the *measured* fixation locations. For example, figure 5.1 depicts a saliency matrix. If we measure fixation locations in points like A, then this saliency matrix is "effective", while if B is a typical measured gaze direction, then this saliency map is not effective at all.

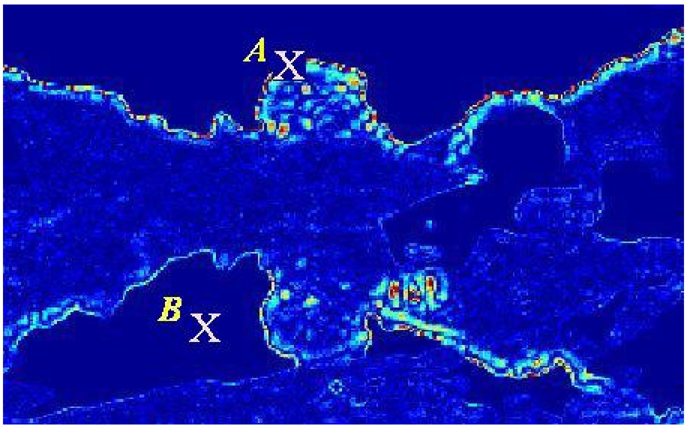

Figure 5.1: Receptive field "efficiency": if we measure fixation locations in points like A, then this saliency matrix is "effective", while if B is a typical measured gaze direction, then this saliency map is not effective at all.

5.1 Saliency calculation: preparing the 40 receptive fields

For calculating saliency values I have used receptive fields (RF), as a biologically inspired solution. Its' schematic structure can be seen on figure 2.3, which I have approximated according to figure 5.2. The smallest was one pixel for the central part surrounded by a one-pixel-width belt. This matched for 0.5°. The largest (the 40th) has had a 79 pixel caliber central region surrounded by a 39 pixel belt.

The parameters depicted on figure 5.2 have been calculated as follows: (x is a simple index, which corresponds to the size. It takes values from 1 to 40.)

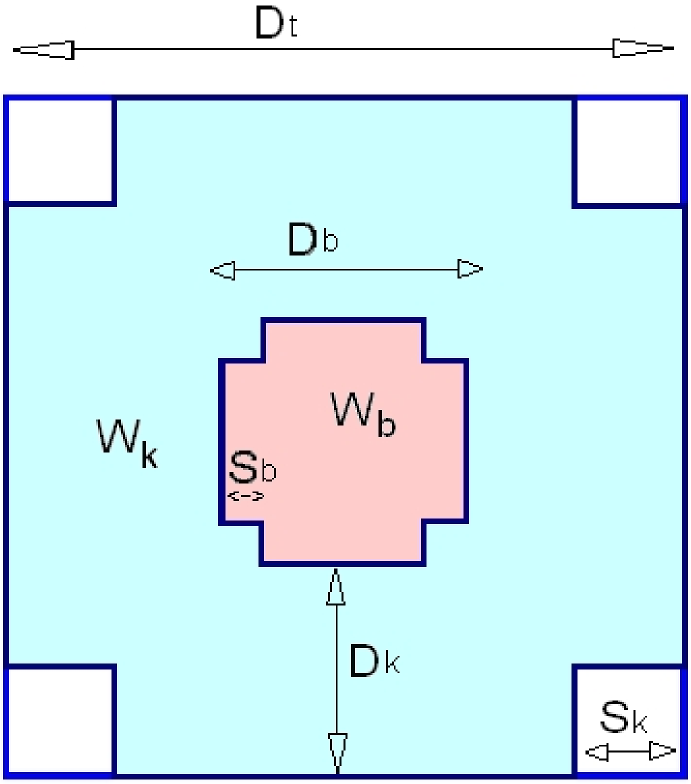

Figure 5.2: The approximation of a general receptive field (RF) [8]. Circles have been approximated with squares chopped down on their corners. The main principles have been: 1) to keep the neuromorphic ratios: In degree, half of the RF should belong to the central part and half to the surroundings. 2) different RF sizes should return the same saliency value if they receive their optimal stimulus, and 3) the saliency value should be zero if the entire RF is exposed by homogeneous light. For precise values see the text.

$D_t(= d_x) = 4x - 3$ the length of the outer square side in pixels

$D_b = 2x - 1$ the length of the inner square-side in pixels, that is the central part's square-side

$D_k = x - 1$ the width of the surrounding ring in pixels

$S_b = round\left(\frac{D_b}{6}\right)$ the length of the square-side that was cut off from the inner square in pixels

$S_k = round\left(\frac{D_t}{6}\right)$ the length of the square-side that was cut off from the outer square in pixels

$N_b = D_b^2 - 4S_b^2$ the number of the pixels in the central part

$N_k = D_t^2 - 4S_k^2 - N_b$ the number of the pixels in the surrounding region

$W_k = \frac{-A}{255*N_k}$ the weight of the surrounding part; this is necessary for the saliency calculation

$W_b = \frac{A}{255*N_b}$ the weight of the inner part; this is also necessary for the saliency calculation

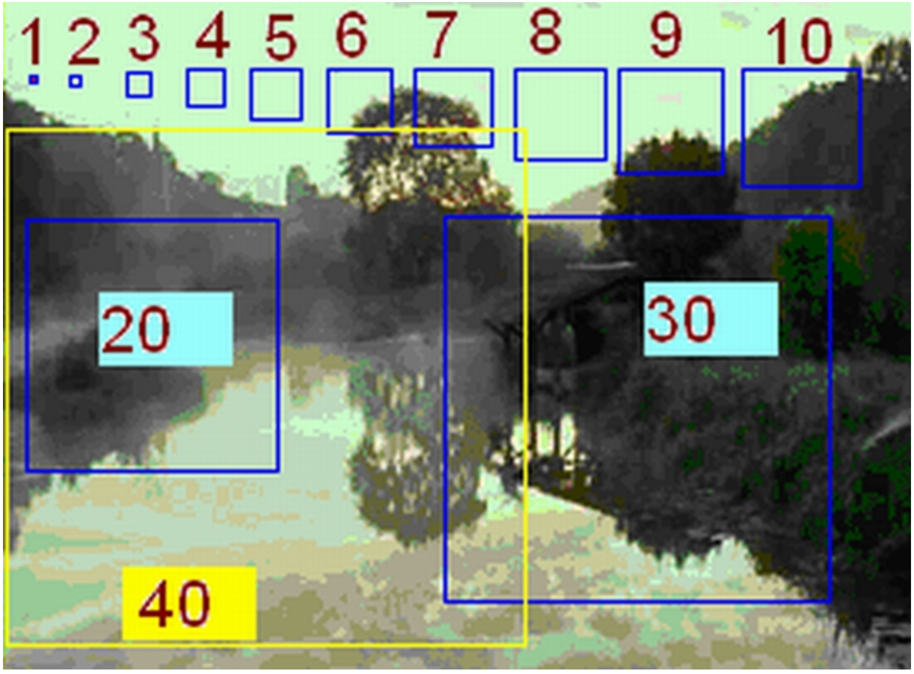

Figure 5.3: The 40 different receptive field sizes I have investigated, spreading from 0.5° up to approximately 26°, expressed in terms of the viewing angle. The squares indicate the diameter of the given RF, which is the "d_x" in the above equations. $x = 1$ is the smallest, $x = 40$ is the largest. The RF sizes increase linearly with respect to the index 'x'. The exact relation between the α viewing angle and the x index is $\tan\frac{\alpha}{2} = \frac{4x-3}{100}0.147$.

'*A*' is an arbitrarily chosen scaling parameter, corresponding to the maximum saliency value that an RF can return if it receives its' optimal stimulus. (A stimulus is *'optimal'*, if both the central and the surrounding part of the RF get the stimuli they respond to with the higher intensity. For example, light appears in the central part and disappears in the surrounding area of an ON-center Off-surrounding RF).

For the sake of accuracy, I note that the index $x = 1$ is an exception from the prior formulas. It corresponds to a one pixel centered, eight pixels surrounding receptive field.

The exact relation between the α viewing angle and the x index – due to the resolution and monitor size – is: $\tan\frac{\alpha}{2} = \frac{4x-3}{100} * 0.147$.

5.2 The rough course of the calculation

- KNOWN are the
  - frames of the training video set,
  - a set of measured fixation locations recorded on this video,
  - the ten retina channel output for arbitrary input,
  - the 40 different RFs, and how the saliency is calculated from an RF and from a retina-channel output. (by convolution)
- Determining the output of the ten retina channels, for all the frames of the training video (– that is 10 matrix/frame).
- Determining the saliency maps according to the 40 RFs, for all the retina channel outputs (– that is $10 * 40 = 400$ matrix/frame).
- Defining the saliency values in the in the measured fixation locations for all channels, all the RF sizes (– that is: every fixation-location – frame pair entails the calculation of 400 saliency values.)
- Based on the above data, producing the functions representing the *average saliency value at the attendant locations* in the function of the *RF size*, for all the ten channels (figure 5.4).

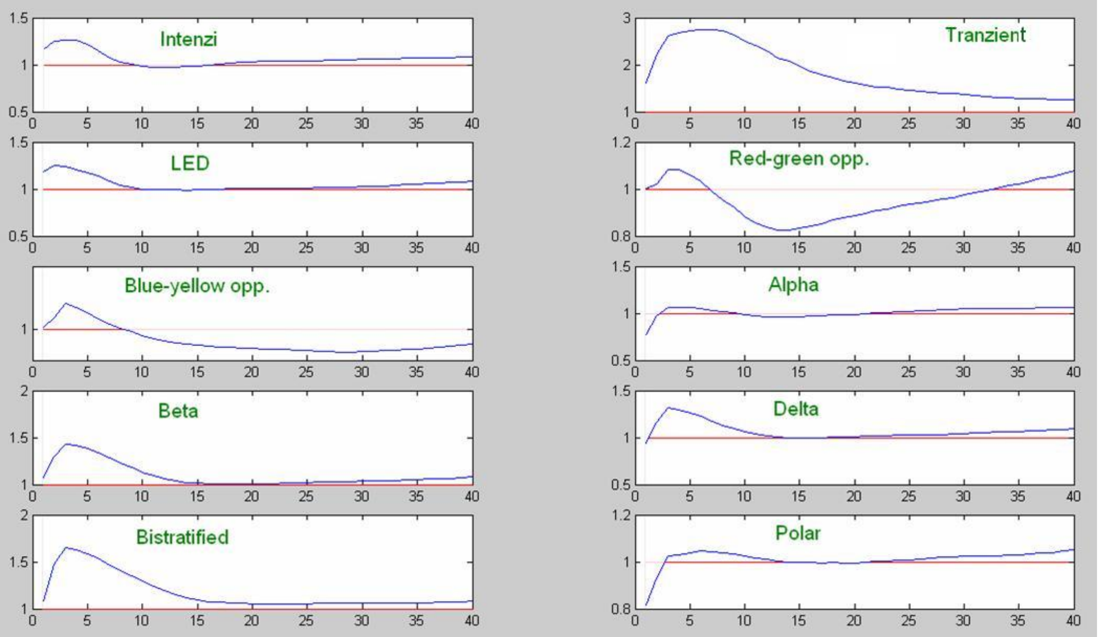

Figure 5.4: The average saliency value at the attendant locations in the function of the RF size, for all the ten channels. The first approximation for the the 'optimal' receptive field sizes: where the curves reach their maximum.

Although it will be always indicated, in this point I note that throughout the dissertation, the channel-enumeration will always follow the next order:

1	Intensity
2	Transient
3	Local Edge Detector (LED)
4	Red-green opposition
5	Blue-yellow opposition
6	Alpha
7	Beta
8	Delta
9	Bistratified
10	Polar

5.3 The detailed course of the calculation

From an engineering viewpoint, receptive fields are *filters* applied on the different retina channels. A RF is determined by a $[d_r \times d_r]$ matrix and the images are defined as [M $\times N$] sized matrices, denoted as $IM_{k,c}$ (these are the outputs of the ten retina channels) for every frame of the input video. The notations I have used are the follows:

R	the number of the receptive field types (sizes); I have worked with R=40.
r	(actual) receptive field type, $r \in \{1, 2, \ldots, R\}$
d_r	the size of the 'r' receptive field, $d = 4r - 3$.
(x, y)	the coordinate of the measured gaze direction (every coordinate-pair belongs to a k frame on which it was measured, where the saccade ended)
M, N	the size of the input video; M: width, N: height; $M = 273$, $N = 201$
C	the number of the channels, C=10
c	(actual) channel number, $c \in \{1, 2, ..., C\}$
K	the number of the frames, $K = 267$
k	(actual) frame number, $k \in \{1, 2, ..., K\}$

The IM image matrices contain values between -127 and +128. -127 is black, 128 is white and 0 is middle-grey. This shifting (compared to the conventional bitmap valuing) is due to the antagonistic behavior of the receptive field weights: the outer part of the receptive fields has negative weight, whereas the inner part has positive. Thus a stimulus, which "fits" to a receptive field, gives a maximal (absolute) value: a big positive, if it fits an ON-centered OFF-surrounded RF, and a big negative ("big" in absolute value) in the other case.

Before the measurements, I have calculated the SM saliency matrices for all the IM image matrices by all the R=40 receptive fields: this means $K \times C \times R$ saliency matrices, that is 400 for every input frame. Thus the saliency matrix is different for every frame, every channel and RF type:

$$SM = SM_{k,c,r} = IM_{k,c} * RF_r \tag{5.1}$$

where '*' denotes convolution. With CNN terminology, RFs are 'A' templates. I have defined the SM saliency map as an $[M \times N]$ matrix, so the outer 'ring' of the matrix has been cut off, which has become into existence because of the convolution.

Let us define the (i, j) coordinates of the (x, y) centered, d sized region in an arbitrary matrix as follows (these are simply those matrix indexes that belong to the d sized region of the $x - y$ point):

$$S_{(x,y,d)} = \left\{ (i,j) \mid \max_{\substack{1 \le i \le M \\ 1 \le j \le N}} \{|i-x|, |j-y|\} \le d/2 \right\} \tag{5.2}$$

During the measurements, I have recorded the *(x,y)* coordinate pairs (the fixation locations) and the corresponding k frame-number for each *(x,y)* fixation location, then I have allocated the proper SVM saliency matrix-segments for all these data-triplets:

$SVM^{k,c,r}_{(x,y)}$ is the $[d_r \times d_r]$ sized, (x,y) centered segment of the $SM^{k,c,r}$ saliency matrix.

(This means $C \times R = 10 \times 40 = 400$ matrix segments for every measured [(x,y), k] triplet.)

In order to define the final saliency value in a given location, I have put a discrete Gauss-filter with the same size and position with the receptive field. The next step has been the creation of these filters in all the $R = 40$ sizes, the discrete form of (5.3), where t is the radius and σ is the standard deviation.

$$G_t = \frac{1}{\sqrt{2\pi\sigma}} e^{-t^2/2\sigma^2} \tag{5.3}$$

For example, the 3×3 discrete Gauss filter is

$$G_1 = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} /16 \tag{5.4}$$

(I note that the bigger ones can be obtained, for example, with repeated convolution from G_1.)
In CNN terminology these filters also act as ‘A’ templates.
With these arrangements we can assign a scalar value for every measured [(x,y), k] data-triplets as follows:

$$SalVal_{(x,y)}^{k,c,r} = \frac{\sum^{M}\sum^{N}\left(SVM_{(x,y)}^{k,c,r} * G_r\right)}{\overline{SM^{k,c,r}}} \tag{5.5}$$

where $\overline{SM^{k,c,r}}$ is the average value of the $SM^{k,c,r}$ matrix and G_r is the discrete Gaussian matrix whose size is also d_r like as in the rth receptive fields.

‘*’ can be interpreted as a filter or convolution. In the latter case, again, I “cut off” the outer ring of the result matrix, which happens to be there because of the convolution.
Hence the normalized saliency value of the cth channel, rth receptive field size, belonging to the kth frame, where the measured fixation location was *(x,y)*, is defined by equation (5.5). For adjusting the model’s free parameters these values were essential.

Let P denote the number of all the measured (and used) fixations. (Some of the measured fixation locations “fell out” from the processing, for example because they followed a saccade less then 1 degree.) With these, the average saliency value arising on channel c by the rth receptive field is

$$\overline{SalVal_{c,r}} = \frac{\sum^{P} SalVal_{(x,y)}^{k,c,r}}{P} \tag{5.6}$$

To find those r^* RF sizes for every c channel where the relative saliency values reach their maximum, I have defined:

$$r_c^* = argmax_r\left\{\overline{SalVal_{c,r}}\right\} \tag{5.7}$$

These are those “*optimal*” receptive field sizes, which had to be defined.

The above deduction shows how the saliency values, and through them, the optimal RFs are determined. However, this train of thought does not take into account, that the different saccades are evoked by different channels. In the next chapter I investigate different assumptions targeting the question of these channel-weights.

Chapter 6

The contributing channels - different approaches

Probably everybody has a share in the experience when a cat unexpectedly glides away next to him, in a dark street. Although this is just a small change in the entire visual environment, it evokes a strong inducement to look there. In this case, it is the *motion* that provokes the attentional shift towards the cat. In the same time, when one has a look at a *steady* picture (that is, motion can not play any role in directing the gaze), the observed locations are not random, but they follow a well defined pattern as well [53, 63] – determined by different channels. In the present model, it is crucial to select the contributing channels properly, together with the *extent* of their contributions (– which are called "weights" in engineering systems).

During this process I strongly exploit the widely accepted 'principle' of attention research, that *saliency attracts visual attention*, primarily under bottom-up conditions [6, 14, 62, 64, 65, 66, 67, 68, 69].

In this chapter I discuss three different approaches targeting the questions of

- which are those channels that contribute in provoking a saccade?
- (related question: *how many* out of the ten channels "work" at a time?)
- define the proper extent of these contributions, that is, to define the *weights* of the channels in question
- what are the corresponding optimal RF sizes for the individual channels

The first two approaches are based on the phenomenon that characterizes attentional mechanism in almost all levels, namely that stimuli *compete* with each other for attention. To realize this, I have treated those channels as saccade generators, to which the highest saliency values belong at the attendant location, by *any* RF size. In other words, only one outstandingly high saliency value being created by a given channel by a given RF size is enough for the characterization of this channel as a saccade-triggering one. The first two approaches are based on this phenomenon. The corresponding curves look very similar, according to the expectations.
To achieve this, in the first approach (as it is detailed in section 6.1) I have set a *threshold* with the aim of selecting the 'appropriate' (that is: saccade-generator) channels. The threshold has been a given percentage of the maximal saliency value that has come into existence on the given frame, on the corresponding measured fixation location. I discuss the results belonging to the thresholds of 95% and 75%.
In the second approach (section 6.2) I have *graded* the channels according to their *highest* saliency value and have taken into account only the first few channels (1, 3 and 5, respectively).
In the context of the third approach, I have assumed that those channels take part in the provocation of a saccade, which are salient *in average* on the actual stimuli. The biological background of this approach is that every channel has a big range of different sized RFs, but their distribution could differ strongly. To define the mean saliency, I have determined the saliency values according to all the 40 RF sizes at the measured attendant locations and have averaged it. Section 6.3 details the corresponding results.

Regarding the *estimations* of the channel weights and the optimal RF sizes, the principles are the same in the different approaches. Accordingly, the *channel weights* are proportional with the ratio determining how many times they have been interpreted as saccade-triggering ones, whereas the estimation of the optimal RF sizes follows the calculation described in section 5.3, with the restriction, that only those channels' data have been taken into account, which have satisfied the condition of the given approach (that is, how many times they have been

considered as saccade-triggering ones, during the given process).
These results are visualized on diagram-assembles, similar to figure 5.4, each of whom consists of ten curves – one for each channel. Consequently, these show the average saliency values in the function of the receptive field sizes in the measured fixation locations, recorded on human observers. The "optimal" RF size belonging to the different channels are those, where the corresponding curve reaches its' maximum. Since the different approaches aim to select those channels that generate the several saccades, these curves differ from each other because – according to the different assumptions – different channels have been considered as saccade-triggering ones.

The depicted results belong to the frames where the given saccades presumably *evoked* in the positions where the saccades in question ended. Accordingly, I have used the *preceding* frame compared with the one that I have measured the saccade-end location on. I have applied 8 fps video, meaning 125 ms retrace in time, which is a good approximation for the period between the saccade initialization and the fixation on the saccade-end location.

Firstly I provide the results in sections 6.1, 6.2 and 6.3 (one section for each approach), then, in section 6.4, I discuss the results.

6.1 First approach

6.1.1 Main steps

- The first step is the same in all the different approaches:
  For all the frames with valid fixation location on them, I have defined the saliency values in the given point, for all the ten channels, all the 40 RF sizes – that is: 400 values for all of these frame–location pairs. For an illustration see table 6.1. (A measured fixation location is "valid", if it is the first recorded location after an at least 1° wide saccade, see Appendix B.3.)

	RF_1	RF_2	RF_3	$\cdots$	RF_{39}	RF_{40}
$Channel_1$	614	729	998	$\cdots$	456	312
$Channel_2$	503	402	378	$\cdots$	716	652
$Channel_3$	424	642	768	$\cdots$	1000	861
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$
$Channel_9$	871	956	804	$\cdots$	502	432
$Channel_{10}$	575	842	615	$\cdots$	307	211

Table 6.1: An illustration aiming to help to understand how the calculations have been made throughout the three approaches described hereinafter. Each row corresponds to a channel (there are ten of these), and 40 RF sizes belong to all of them – these are the columns. Once a "valid" fixation location had been measured on a frame, all the 400 saliency values have been determined, according to the 10 channels' 40 RF size. This can be imagined as if a table like this would had been filled out, for all these measured locations. This step is common in the three approaches, whereas the assumptions aiming to select the saccade-triggering channels are different.

- From this step, the procedure pertains to this approach alone:

  I have searched for the largest value among these 400 values (1000 in the example depicted on table 6.1, belonging to the third channel, 39th receptive field, marked with dark background).

- Here, the ASSUMPTION has been, that *those channels trigger a saccade on a given stimulus (input frame), which have at least one RF that produces saliency value reaching a given percent of the maximal value.*

  For example, if this "given percent" is 95%, then the 1st and the 9th channels contribute in repositing the fixation location in the example shown on table 6.1 – besides, of course the 3rd channel, which gives the maximum value.

SOME REMARKS:

- the maximal value is always different (or if it is not, then it is just by coincidence)

- the number of the channels satisfying the condition varies from frame to frame, as well
- the calculations had been completed with different percentages

6.1.2 Experimental results by thresholds set to 95% and 75%

Figures 6.1 and 6.2 belong to the stricter case in which the threshold is set to be 95%, that is, a channel's highest saliency value has to reach the 95% of the *maximal* saliency value (came into existance by *any* channel, *any* RF size) to be taken into account. The results regarding the "efficiency" of the different RF sizes are depicted on figure 6.1, whereas the ones regarding the channel weights can be seen on figure 6.2.

In the case to which figures 6.3 and 6.4 belong to, the condition is not as strict as in the previous one: a channel's data has been taken into account, if it had at least one RF that provided saliency value exceeding the 75% of the maximal value, instead of 95%. That means, that more channels' data have been included during the generation of the curves depicted on figure 6.4.

Regarding the corresponding channel weights, as it can be seen on figure 6.3, the differences are reduced compared with the previous one (figure 6.2), which indicates that the *one* most salient channel very frequently proved to be the Transient one. Even so, the leading role of the Transient channel in guiding visual attention on moving stimuli, still stands out.

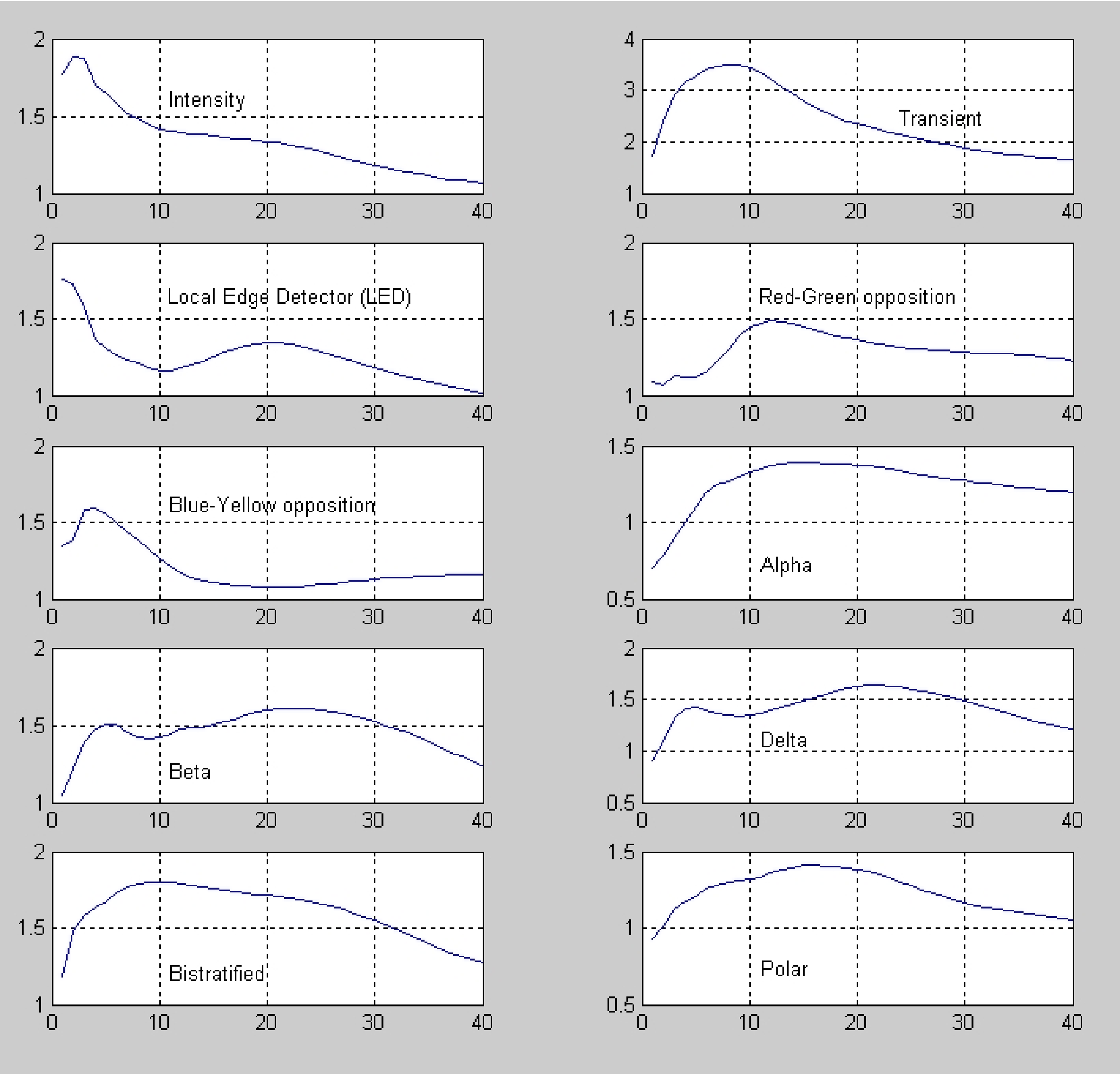

Figure 6.1: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. The corresponding channel names are indicated on the figures. In this case the threshold is set to be 95%, that is, for every [measured gaze direction-frame] pair, only those channel-data have been taken into account, which had at least one RF that provided a saliency value reaching the 95% of the maximal saliency value.

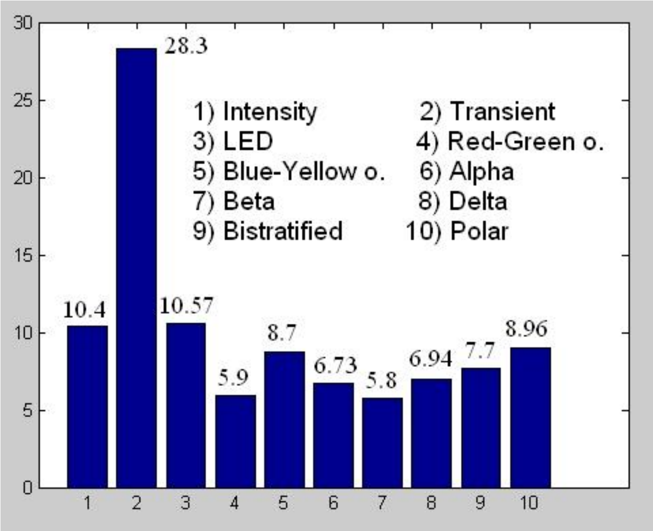

Figure 6.2: Channel weights (that is, the percentages expressing how often they have been considered as the ones generating the saccades) corresponding to the first approach, with the threshold set to be 95%. The outstanding (second) bar belongs to the Transient channel, which responds to motion. The exact values are depicted on the top of each column, in percentage.

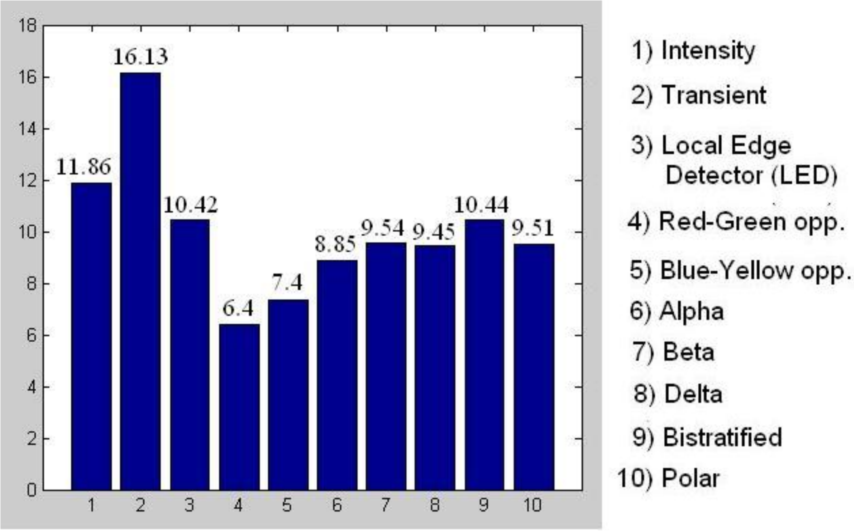

Figure 6.3: Channel weights corresponding to the first approach, with the threshold set to be 75%. The exact values are depicted on the top of each column, in percentage. By the less strict condition the differences among the channels have decreased.

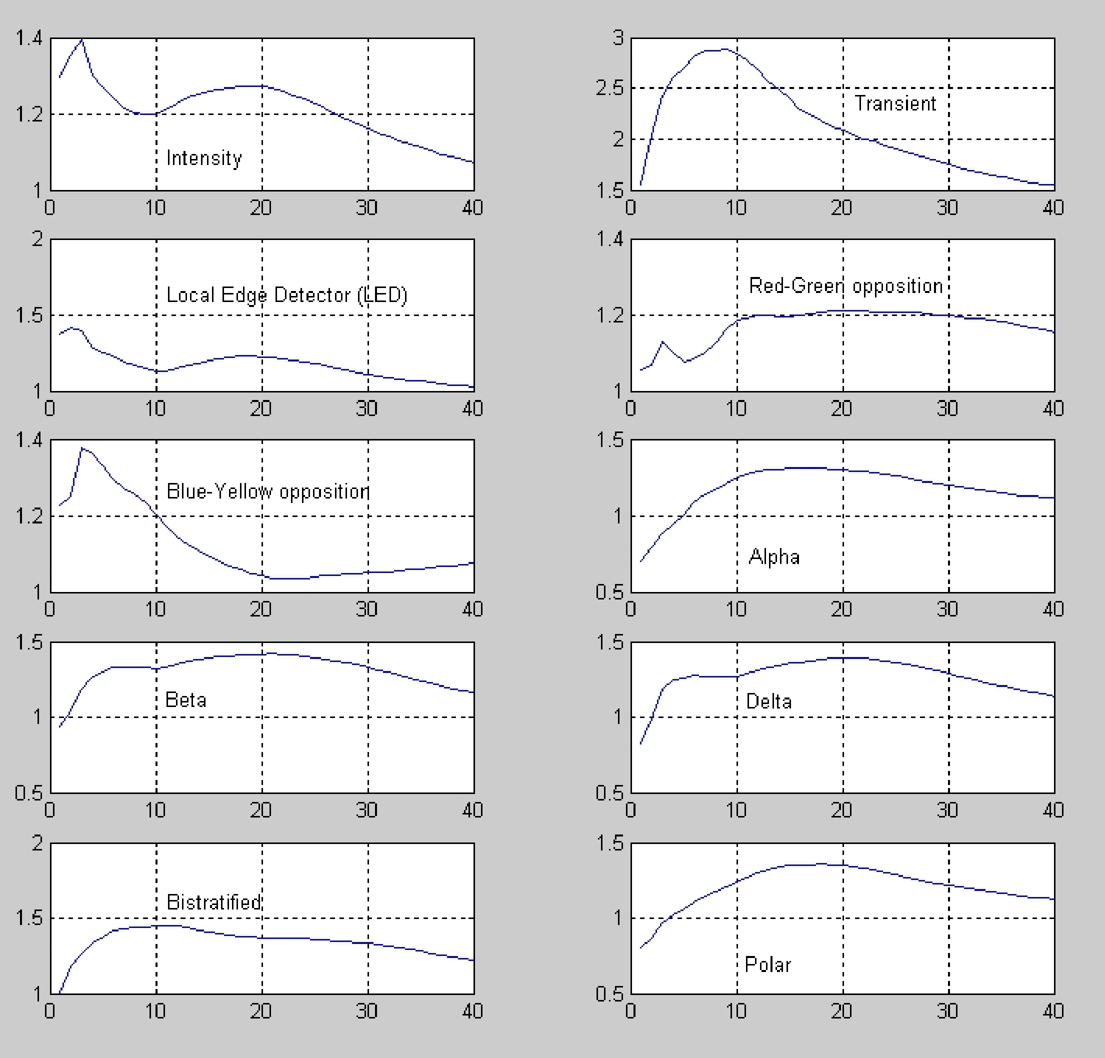

Figure 6.4: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. The corresponding channel names are indicated on the figures. In this case the threshold is set to be 75%, that is, for every [measured gaze direction-frame] pair, those channel-data have been taken into account, which had at least one RF that provided saliency value reaching the 75% of the maximal saliency value.

6.2 Second approach

6.2.1 Main steps

- The first step – which is the same for all the three strategies – is the following:
  For all the frames with valid measured fixation locations, I have defined the saliency values in the given point, for all the ten channels, all the 40 RF sizes – that is: 400 values for all of these frame–location pairs. As an illustration see table 6.1.

- Then I have *ranked* the channels according to their highest saliency values. In the example used previously (table 6.1), assuming that the values are smaller in the missing, dotted parts, the channels in order would be:

  1. Channel 3 (with maximum saliency value: 1000, by RF 39)
  2. Channel 1 (with maximum saliency value: 998, by RF 3)
  3. Channel 9 (with maximum saliency value: 956, by RF 2)
  4. Channel 10 (with maximum saliency value: 842, by RF 2)
  5. Channel 2 (with maximum saliency value: 716, by RF 39), etc.

- Here the ASSUMPTION has been, that *the first few most salient channels trigger the saccades, according to the above ranking.*

For example, if these "first few" are two, then the 3rd and the 1st channels contribute in generating the next saccade. As it follows from the algorithm – in contrast with the previous approach – in this case the number of the participating channels are always the same.

6.2.2 Experimental results for the first 1, 3 and 5 most salient channels

Figures 6.5 and 6.6 record the results that belong to the case when only the first most salient channel's data has been taken into account, the next two diagrams,

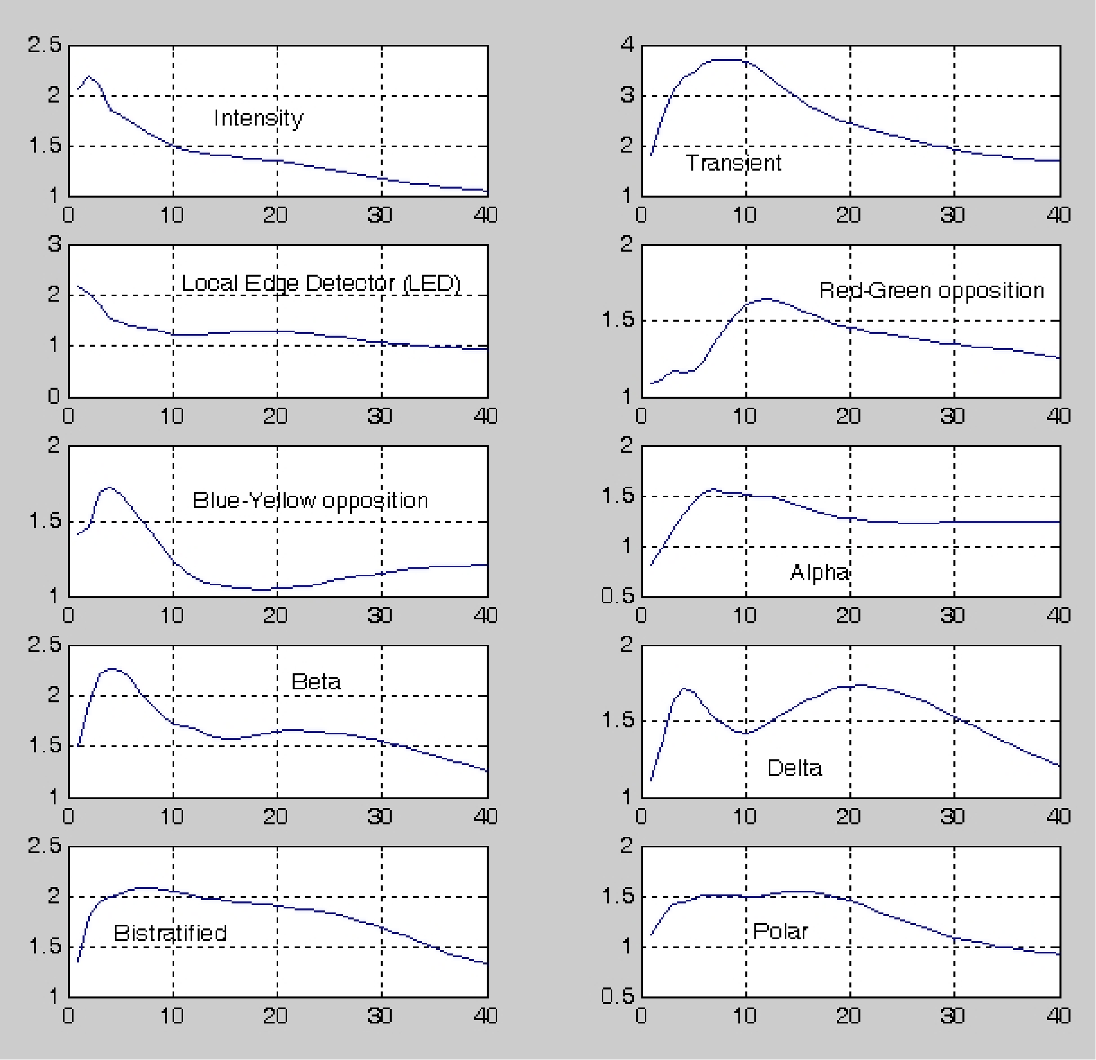

Figure 6.5: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. The corresponding channel names are indicated on the figures. In this case only those channel's data has been taken into account, to which the highest saliency value belonged to – that is: one channel for each measured fixation location.

figures 6.7 and 6.8 belong to the case in which the first 3 most salient channels have been treated as the ones generating the saccades, and finally, figures 6.9 and 6.10 belong to the case of the first 5 most salient channels.

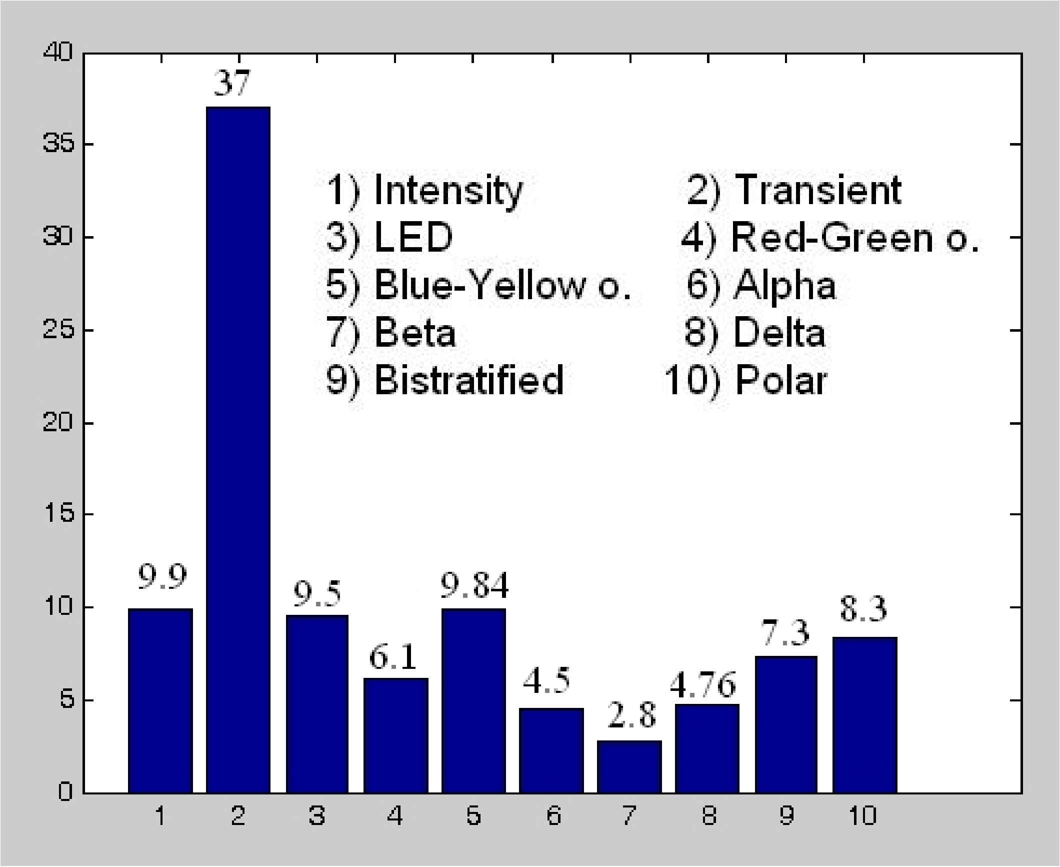

Figure 6.6: Channel weights corresponding to the second approach. In this case only one channel's datum (the one that provided the biggest saliency value) has been taken into account for all the measured fixation locations.

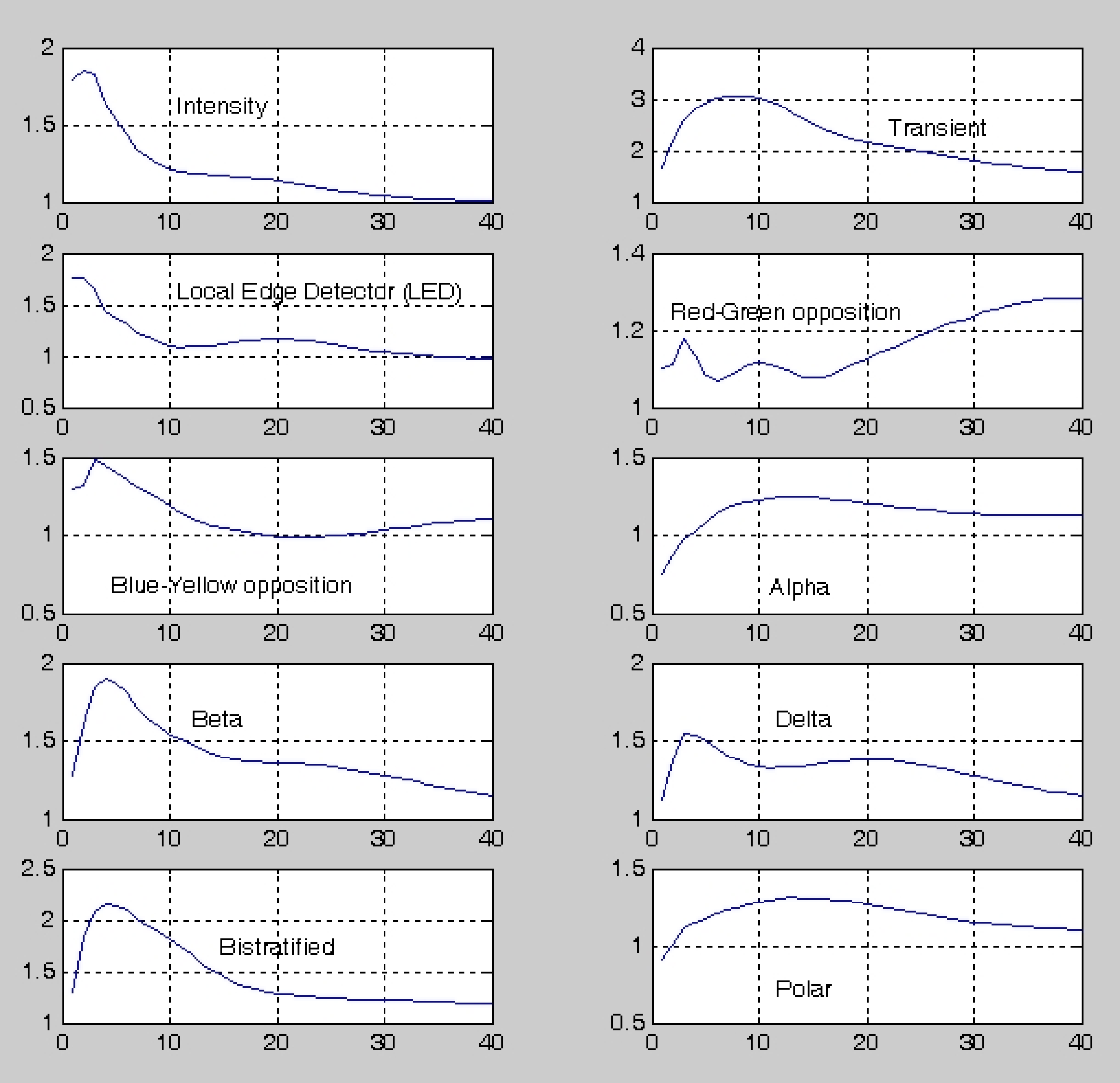

Figure 6.7: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. In this approach, the channels have been ranked according to their highest saliency values. The above diagram belongs to the case in which the first 3 channels' data have been taken into account.

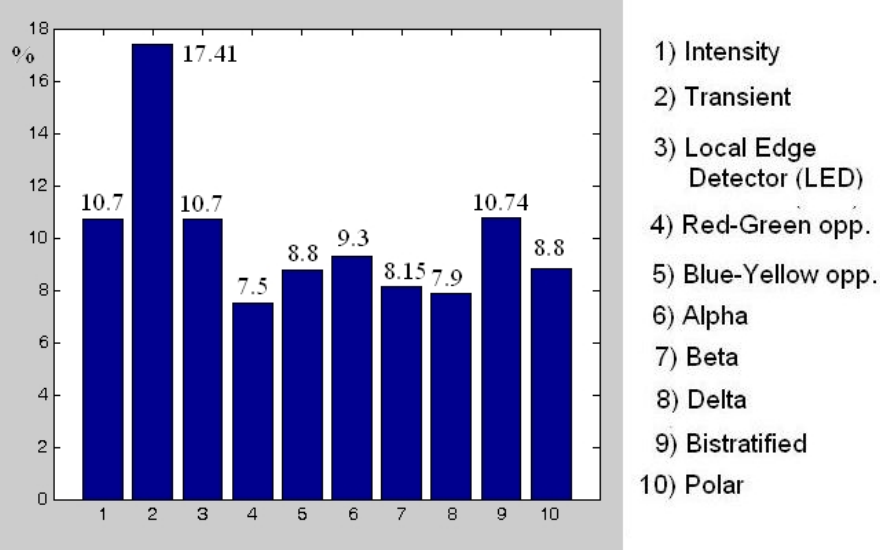

Figure 6.8: Channel weights corresponding to the second approach, when the first 3 most salient channels' data have been taken into account. These "weights" are the percentages expressing how often they have been considered as the ones generating the saccades, that is, how often they have satisfied the criteria of the present approach.

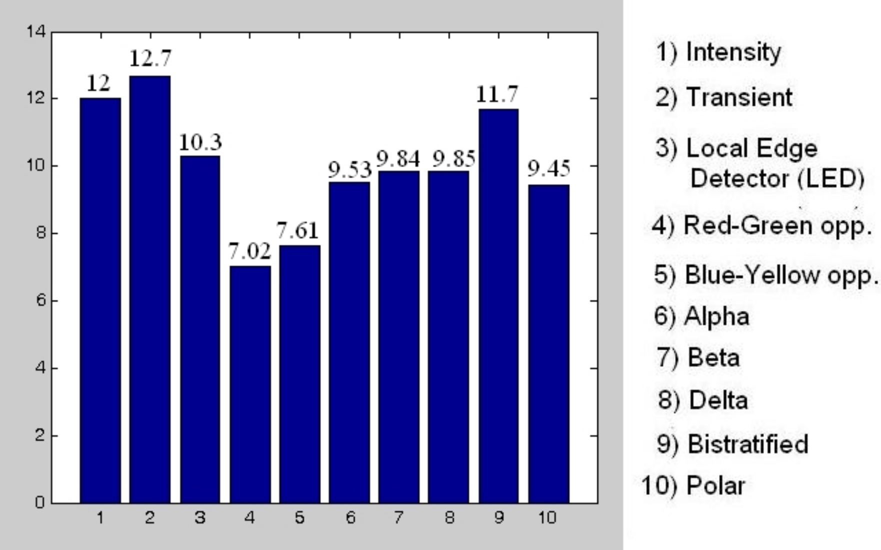

Figure 6.9: Channel weights corresponding to the second approach. The first 5 most salient channels' data have been taken into account. The *relative* role of the Transient channel (second column) apparently decreases.

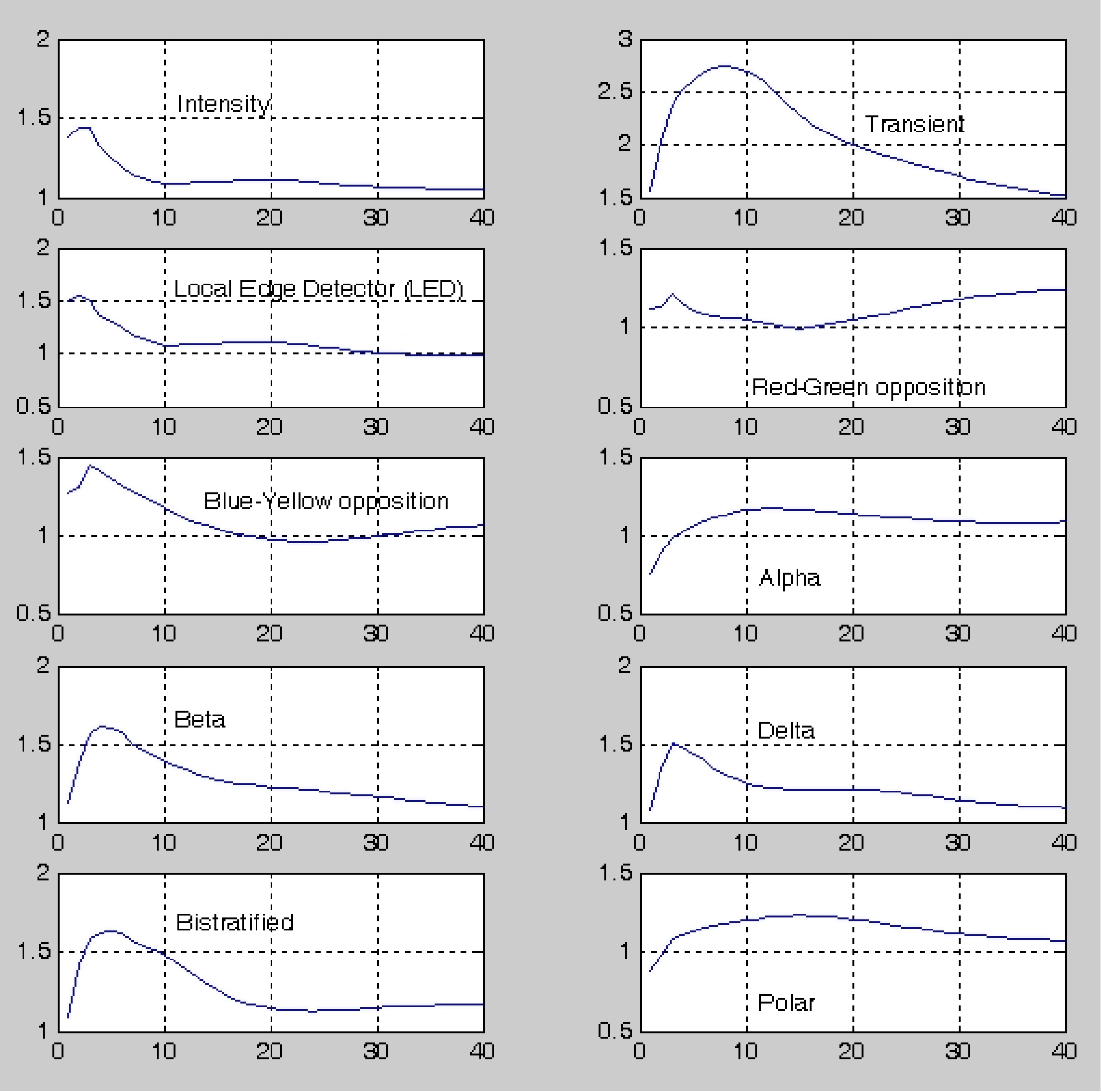

Figure 6.10: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. In this approach the channels have been ranked according to their highest saliency values. The above diagram belongs to the case in which the first 5 channels' data – that is, *half* of the channels – have been taken into account.

6.3 Third approach

In this approach I have selected the channels to which the highest *average* saliency values belonged to (defined by the 40 RFs), instead of selecting the channels that contained the highest one or two values. According to the expectations, the results belonging to this approach differ more from the first two ones, than those differing from each other. Thus the process described in this section gives new “optimal receptive field sizes” and corresponding channel weights (table 6.3) than to the ones belonging to the first two approaches (table 6.2), which could have been bracketed because of their similarities.

6.3.1 Main steps

- For all the frames with valid measured fixation locations, I have defined the saliency values in the given point, for all the ten channels, all the 40 RF sizes – that is: 400 values for all of these frame–location pairs. (An illustration can be seen on table 6.1.)
- Then I have defined the average saliency values for all the ten channels, among the 40 RF sizes. On the example depicted on table 6.1, this means the determination of averages of the 10 different rows.
- Finally I have selected the channel(s) bearing the highest average saliency value(s), according to the assumption that these channels trigger the saccades.

6.3.2 Experimental results for the first 1, 3 and 5 most salient channels

As described above, in this approach I have ranked the channels according to their *average* saliency values appeared in the measured fixation locations. Figures 6.11 and 6.12 belong to the case, where only the first most salient channel has been taken into account (– salient on the average), the next two, figures 6.13 and 6.14 to the case in which the first 3, whereas the last two ones, figures 6.15 and 6.16, correspond to the case, where the first 5 (that is, half of the channels) have been considered as saccade generators.

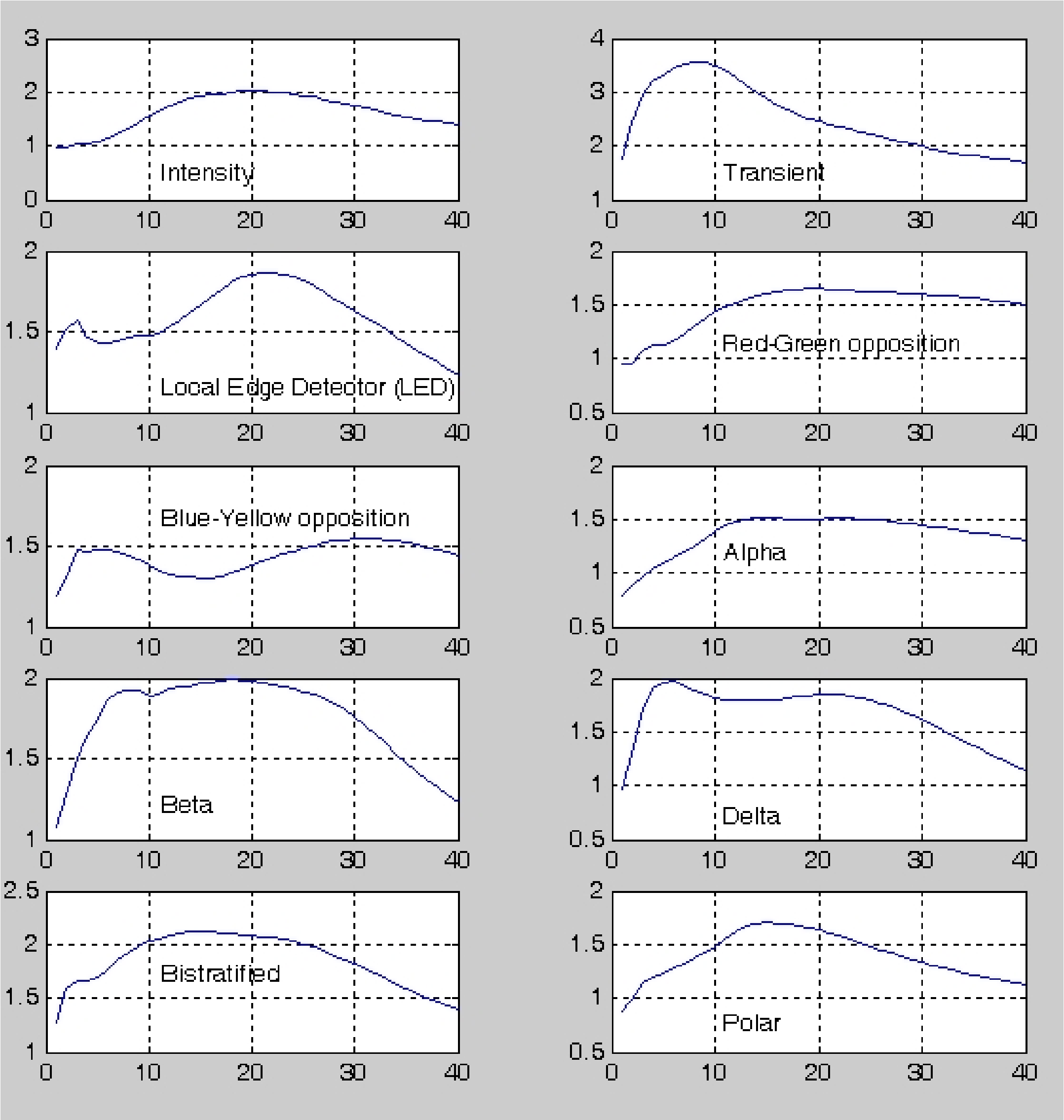

Figure 6.11: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. The above curves include the data of those channels, which proved to be the most salient *in average* at the recorded fixation locations.

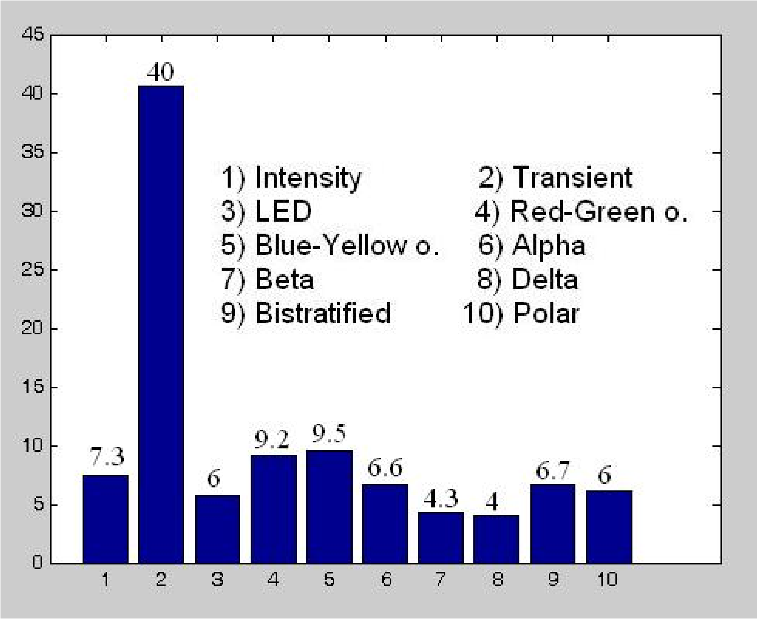

Figure 6.12: Channel weights corresponding to the third approach. The above diagram shows the percentages expressing how often the several channels proved to be the ones with the highest average saliency value, at the recorded fixation locations.

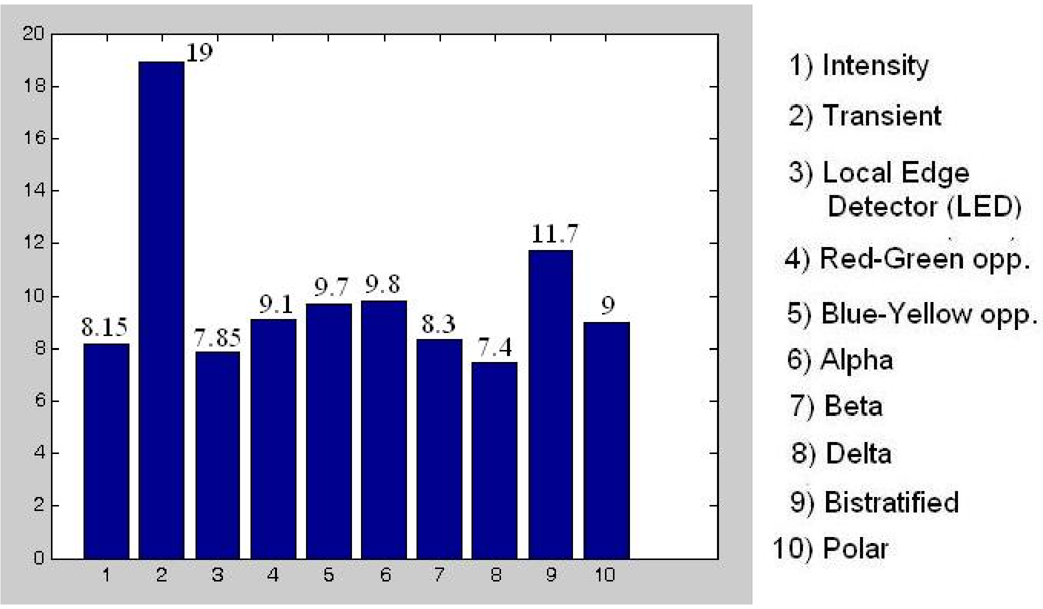

Figure 6.13: Channel weights corresponding to the third approach, namely to the case in which the first 3, in average most salient channels' data have been taken into account.

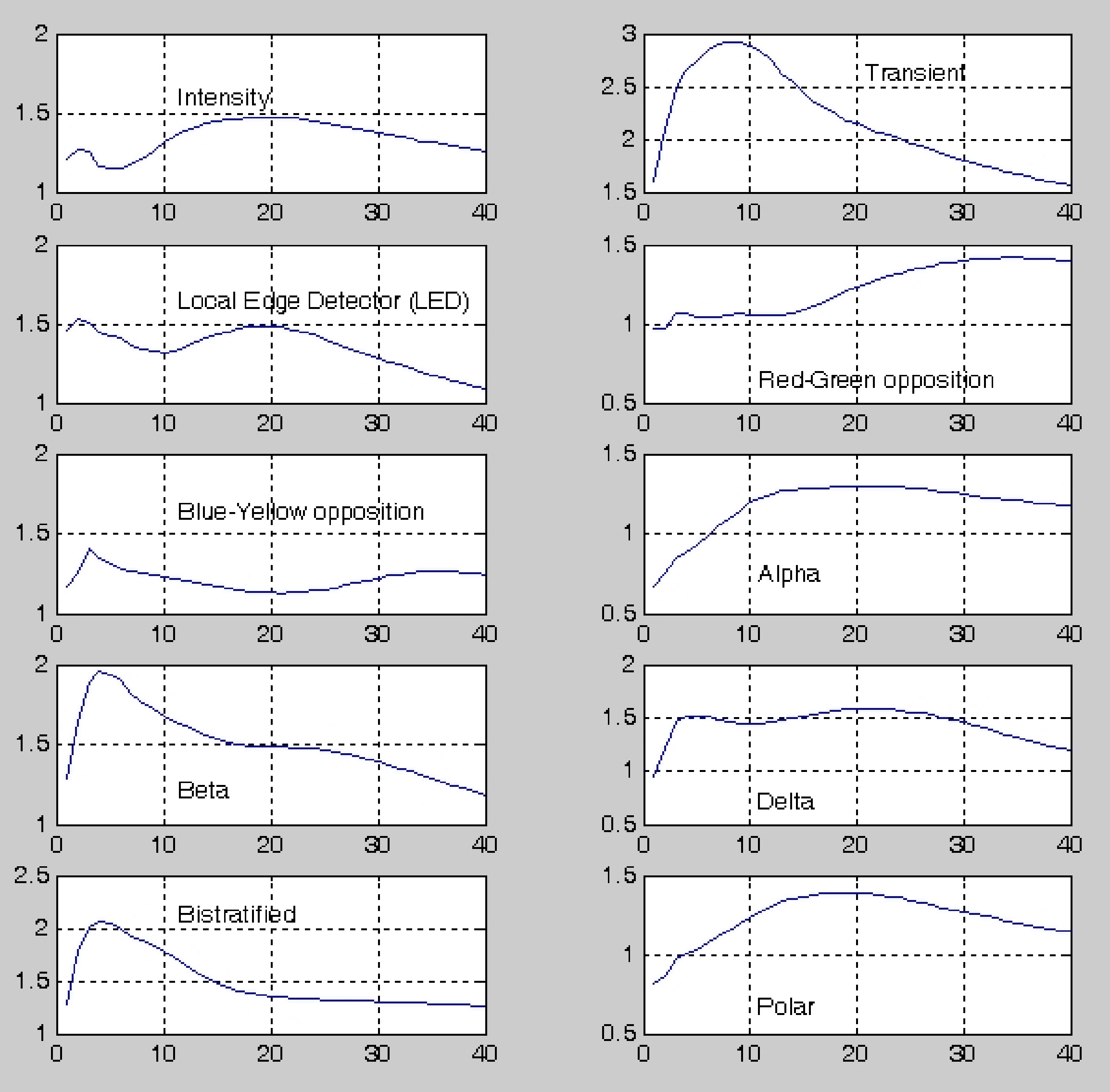

Figure 6.14: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. In this approach I have ranked the channels according to their *average* saliency values, among the 40 RFs, at the attendant fixation locations. The above curve includes the data of the first 3 channels, from this ranking.

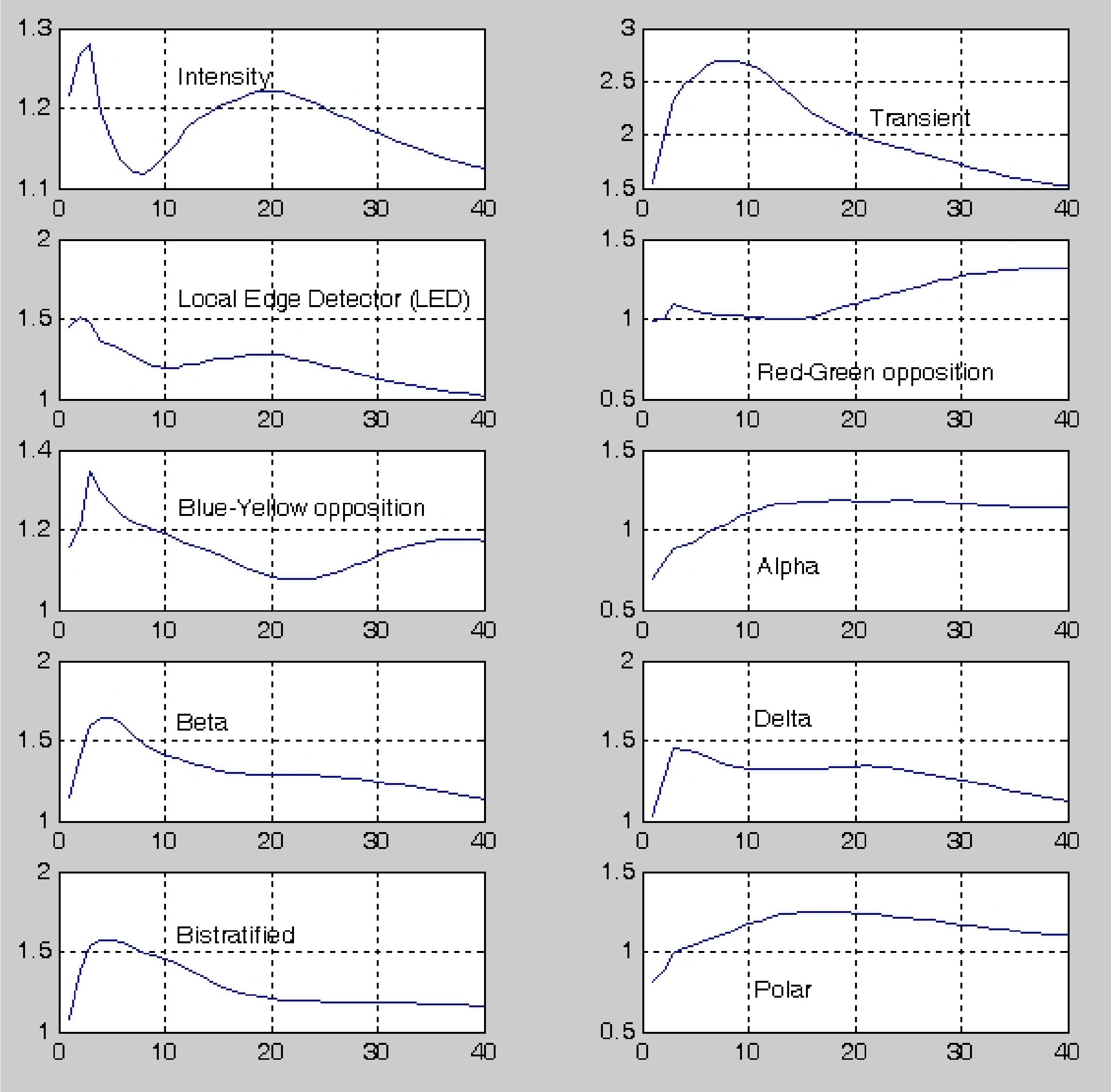

Figure 6.15: The average saliency values at the recorded fixation locations, as a function of the receptive field sizes, for all the ten retina channels, respectively. In this approach I have ranked the channels according to their *average* saliency values, at the attendant fixation locations. The above curve includes the data of the first 5 channels from this ranking.

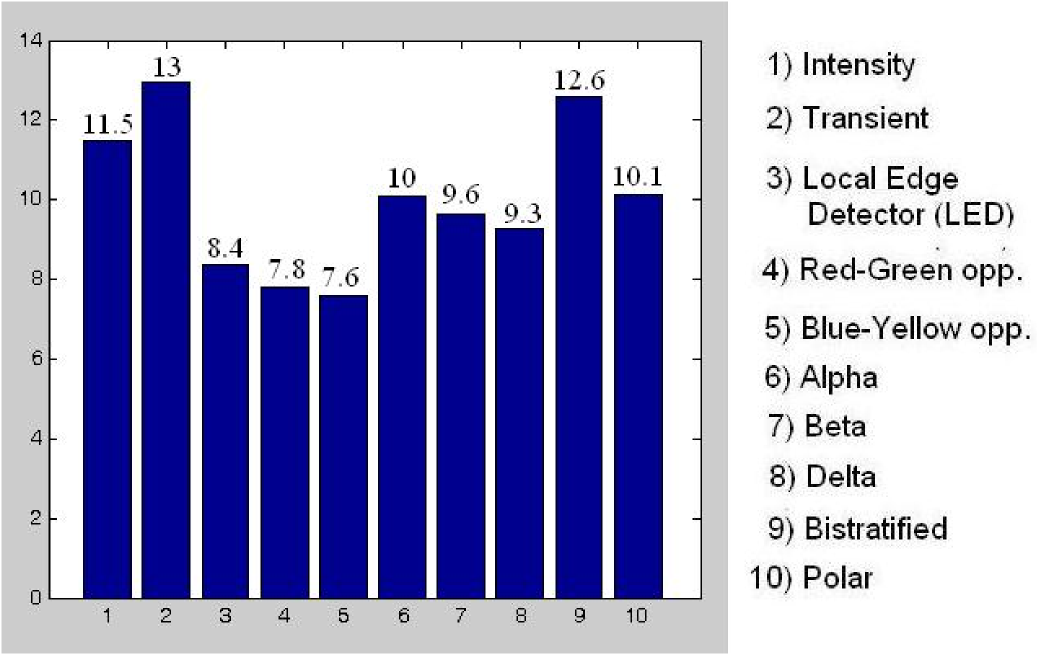

Figure 6.16: Channel weights corresponding to the third approach, namely to the case in which the first 5, an average most salient channels' data have been taken into account – in other words, in this case *half* of the channels are considered as saccade-triggering ones.

6.4 Discussing the results

The prime goal of the measurements and calculations has been to yield an attentional model that selects similar locations to those that a human observer selects, on complex moving natural scenes, under bottom-up conditions. This is not an obvious question, primarily if one considers that two different human observers might easily attend to different locations on the same frame of a video. Moreover, the *same* observer might fixate to *different* locations on the *same* frame, during watching the video for a second or a third time. Even so, there are obviously some criteria that make some point a probable candidate for winning the attention, or oppositely, very unlikely to be attended to [6, 37, 42, 43, 62, 64, 65].
Exploiting these criteria, which are basically the *saliency values* of the several points in the visual environment, artificial systems can be created behaving similarly to living creatures. A basic problem is that these saliency values fundamentally depend on both the *"low level visual features"* by which they are salient or not (color, motion, etc.), and also on the *way of defining* these values in question. Since in the introduced attentional model I use a mammalian multi-channel retina simulator, these "low level visual features"in my case are straight-forward: they are defined by the different channels' behavior (– and thus following a bio-inspired filtering, instead of a heuristic one). But the "way of defining" these saliency values are not that straight-forward: although it is known, that living creatures basically 'apply' receptive fields with this end, which can be mathematically simulated (by a simple convolution), different RF sizes give completely different saliency values. Thus, the fundamental question arises: what sized RFs should be applied on the different retina channels?

Figures depicting the average saliency values in the function of the RF sizes (namely figures 6.1, 6.4, 6.5, 6.7, 6.10, 6.11, 6.14 and 6.15.) target this question. Since they are calculated from data that belong to locations where a human observer *did* attend to during the measurements, these curves reflect the average 'saliency level' for all the channels, created by the different RF sizes, *at the locations that have won the attention.* An artificial model will give the closest results to humans, if it selects fixation locations based on similar saliency values that appear in the living creatures (in the appropriate topographic maps; see sections 2.2.1 and 2.2.2). To ensure this, two fundamental criteria should be satisfied:

1. On a given stimulus, the allocation of the new fixation location should depend on the same low-level visual features (– which in my case means the same channels). The different approaches introduced above, in sections 6.1, 6.2 and 6.3, aim to ensure this criterion.

2. On a given channel, the saliency calculation should result in a similar outcome for the living creature and for the artificial one (– that is, on a given channel the same places should appear salient). In my case, this latter one basically means the selection of the proper RF size for the different channels. These are the "optimal RF sizes", where the different curves reach their maximums.

Before summarizing the results in a table, I make some general comments which have influenced the readings:

- Although the curves belonging to the same channels in the different diagrams show more or less differences (which is innate), the similarities throughout the first two approaches are apparent – primarily from the viewpoint of the maximum-locations. See for example figures 6.1 and 6.5, which belong to different approaches. (Theoretically, since the different approaches select different channels, the resulting curves could be completely different as well, primarily because the different cases include different amount of channels.) During the inspection of the figures please note that the scale division on the vertical axes varies!

  Because of the above similarities, from an engineering viewpoint it is expedient to amalgamate the results of the first two approaches.

- The role of the different channels can be better traced on those cases, where the exact number of the participating channels are traceable (– primarily if this number is 1). In this sense, the second approach gives more accurate results than the first one, hence the readings rely more on this approach (figures 6.5, 6.7 and 6.10) than on the first one (figures 6.1 and 6.4).

- Regarding the channel weights, they – in contrast with the optimal RF sizes – fundamentally depend on the number of the channels (1, 3, or 5) that have been taken into consideration. This is natural, since – for example – if a channel has an indisputable leading role (and the Transient channel indisputably has, see figures 6.2, 6.6 and 6.12), then in those cases when not only the *first* most salient channel's data is taken into account, the *relative weight* of the remaining channels will necessary increase – compare figure 6.3 to 6.2; 6.8 and 6.9 to 6.6, and finally 6.13 and 6.16 to 6.12. (The numbers depicted on the top of the bars are rounding-offs.)
  Since the most independent channel-characteristics can be seen on those diagrams, which take into account *one* channel for every measured fixation location (these are the "first cases" in the different approaches), the corresponding channel weights (figures 6.6 and 6.12) are co-ordinated to the optimal RF sizes. The final results are summarized in tables 6.2 and 6.3.

Some engineering notes:

- Before the measurements there were no clues as to what shaped curves should be expected: would they be smooth with a few optimums, would they show uniform or random distribution, or would they be more like noise, etc.? As it can be seen, most of the curves have one, whereas some of them two cambers. Nevertheless, this does not mean that the corresponding channels unequivocally have exactly two RF sizes. As it has already been mentioned before, the several channels have a wide *range* of different sized RFs, but in unequal amount: one channel favors a given size, another one prefers a different size, etc. These curves reflect much more these 'preferences'.

- Of course, one may consider an other way to define these saliency values as well, which also seems to be accurate: define the saliency values in *every single point* of the visual environment according to all the used RF sizes (40 in my case), for all the "low level visual feature extracting" channels (10 in my case), and then define a weighted average from these values, keeping in mind the above defined curves, which can be interpreted as the *frequency distribution* of the different RFs, for all the channels (see previous bullet). The main problem with this process is not theoretical, but practical: it would demand the calculation of *more than 400 values* (at least in the present set-up) for *every single pixel of every frame* of the input video. In contrast, by choosing an appropriate RF size for each channel, this amount of more than four-hundred values decrease to ten, for all the frames. Since the *speed* and necessary processing demands are also fundamental features for an attentional model, the occasional accuracy that this solution offers is not worth the amount of the additional calculation the process requires.

Keeping in mind the above remarks, table 6.2 summarizes the results belonging to the first two approaches, whereas 6.3 synopses the outcome of the third approach.

Channel name	RF size (in index)	RF size (in viewing angle)	Channel weight (in percentage)
Intensity	3	1.5°	9.9
Transient	9	5.55°	37
LED	2	0.84°	9.5
Red-Green opp.	12	7.57°	6.1
Blue-Yellow opp.	3	1.5°	9.84
Alpha	12	7.57°	4.5
Beta	4	2.18°	2.8
Delta	4	2.18°	4.76
Bistratified	4	2.18°	7.3
Polar	15	9.58°	8.3

Table 6.2: An estimation for the quested parameters highlighted with red question-marks on figure 4.1 (*optimal RF sizes* and *channel weights*) **belonging to the first two approaches**, detailed in section 6.1 and 6.2. The first column shows the channel names, the second column indicates the *index* of the optimal RF size for the different channels, whereas the third column depicts the corresponding RF sizes in *viewing angle.* The fourth column indicates the corresponding channel weights, which are basically the values depicted on figure 6.6. During the validation, the results belonging to this approach will often be referred to as '**M rf**' (- which is only a fancy name used for abbreviation, implying the process, during which an arbitrary RF can be considered as reaching the maximum.)

Channel name	RF size (in index)	RF size (in viewing angle)	Channel weight (in percentage)
Intensity	20	12.9°	7.3
Transient	9	5.55°	40
LED	21	13.6°	6
Red-Green opp.	20	12.9°	9.2
Blue-Yellow opp.	31	20.2°	9.5
Alpha	22	14.2°	6.6
Beta	19	12.2°	4.3
Delta	4	2.18°	4
Bistratified	15	9.6°	6.7
Polar	15	9.6°	6

Table 6.3: An estimation for the quested parameters (*optimal RF sizes* and *channel weights*) **belonging to the third approach**, detailed in section 6.3. The first column shows the channel names, the second column indicates the index of the optimal RF size for the different channels, whereas the third column depicts the corresponding RF sizes in viewing angle. The fourth column indicates the corresponding channel weights, which are basically the values depicted on figure 6.12. During the validation, the results belonging to this approach will often be referred to as '**Avg**' - implying the average calculation.

Chapter 7

Validation

The present chapter summarizes the verification results of the model, that has been adjusted according to the parameters yielded previously. Essentially, there are two main ways in which the model can work, regarding the *creation of the final saliency map* from the channel-based saliency maps (bottom of figure 4.1).

1. One way is to use the channel weights that have been estimated beforehand and summarized in tables 6.2 and 6.3 (third column). That is, a weighted sum of the channel based saliency maps form the final map for every frame, where the weighting is the one mentioned before. This will be referred to as "*constant channel weighting strategy*".

2. An other strategy is to take into consideration the stimuli, that is the frames one wants to make predictions for. In this case the above mentioned weights will not be used, but the channel-proportion will depend on the actual stimulus' features: if it is a red house in a green field, the red-green opposition channel will be an important one, if it is a flying bird in front of a mountain, then the Transient, etc. (Of course, these actual channel weights are not decided beforehand, but they depend on the features of the corresponding saliency maps.) This method will be referred to as "*dynamic (continually updated) channel weighting strategy*".

7.1 verification results on constant channel weights

With the aim of testing how close the models predictions are to human fixations, I have proceeded as follows:

- Firstly, I have produced the final saliency map applying the channel weights estimated beforehand (see sections 6.1 - 6.3), according to both results. (The parameters belonging to the approaches are summarized in tables 6.2 and 6.3, respectively.)

- Then I have made *predictions* for the gaze directions. These were locations (x-y coordinate pairs) which the model has calculated as the most probable fixation locations. This means that if the model and the used assumptions are correct, a human observer will attend to these locations with a higher probability than to other points. There were more of these predicted locations (exactly four) to every frame, ordered by decreasing probability: the first location has been calculated as the most likely fixation location, the second one as the second most probable, and so on. Practically, these probability values were saliency values calculated according to the different approaches.

- In the same time, I have made human gaze direction measurements as well, on the same video that the predictions had been made for, for the purpose of *comparing* the predictions with the measurements.

  - I have defined HIT, as if the distance between the predicted and measured fixation location was less than 5°. (Accordingly, *accidental chance* was the product of an area of a 5° radius circle and the number of predictions (1, 2, 3 or 4) divided by the area of the monitor.)
  - Mathematically, the accidental chance (C_a) has been calculated as follows:

$$C_a = \frac{T_{rh} \cdot N_{pred}}{T_m}$$

where T_{rh} denotes the area of a (in this case) 5° radius circle, N_{pred} is the number of the predictions (1,2,3 or 4), and T_m indicates the sphere of the monitor.

Figure 7.1 shows the results.

Driven by the unambiguously outstanding role of the Transient channel on moving stimuli under bottom-up conditions, I have made an other comparison as well: I was interested in the model's accuracy, if the predictions are made based on the Transient channel *solely.* In other words, the master saliency map – in this case – is equal with the Transient channel-based saliency map. The results are included in figure 7.1, besides the comparison of the two above introduced methods, and the corresponding accidental chance. Other researchers have also observed similar surprising efficiency of this channel, when it is applied alone, under similar conditions [42].
As it can be seen, this method is almost as effective as the other strategies. On the whole, these estimations are quite effective: the first four predictions contain the *measured* fixation location with approximately 70% for arbitrary subject.

This shows, that on moving stimuli, during bottom-up attentional conditions, the commanding role of the Transient channel is undoubted. Nevertheless, in a *general* attention model, by all odds all the channels have their own role (it is enough to think of motionless visual environment).Thus, the whole attentional method has to be under the control of the other channels as well. Moreover, under top-down conditions, during which search being based on complex visual features comes to the front, probably the importance of these channels further increases.

Even so, the best approach is apparently the one using the *averages* (middle bar in the triplets with dots in it). Although the differences are not big, in all the four investigated cases this one proved to be the most efficient, while the other two go 'neck and neck'.

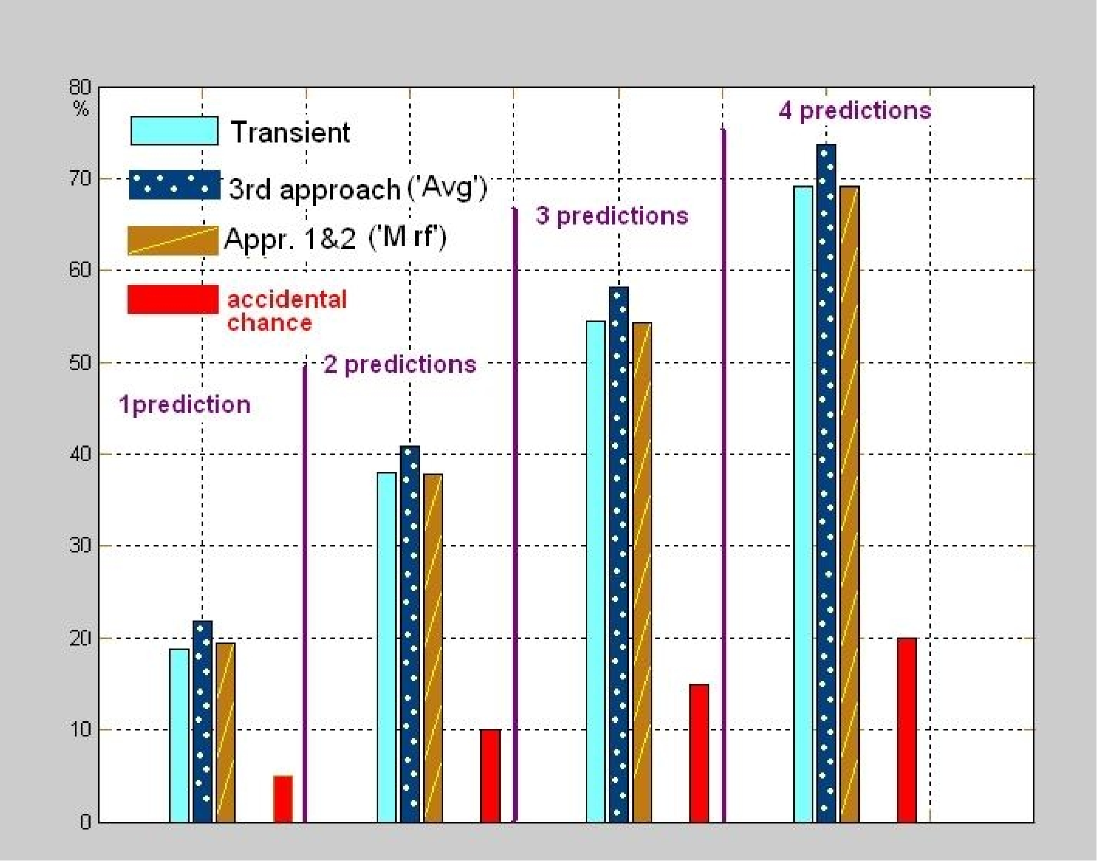

Figure 7.1: Validation results for the constant channel weighting strategies, for 4 predictions/frames. The left-most bar in each triplet shows the results based on the Transient channel's saliency map, the middle one (with dots) belongs to the "third approach" detailed in section 6.3 and summarized in table 6.3, whereas the right ones (with stripes) reflect the outcome of the predictions made by the parameters of the first two approaches (summarized in table 6.2). The red bar on the right of each triplet indicates the accidental chance for making correct predictions under similar conditions.

Here I would like to remark that in order to understand the importance of the parameters' correct estimation, I have also made predictions based on *randomly generated parameters* (channel weights and RF sizes), and compared them to the human gaze direction measurements. In this respect, the correlation percentage between the above mentioned parameters and the measured ones, has been slightly above 30%. This means, that it is better than the accidental chance by around 10%. (According to the above, the major contribution is most probably due to the proper selection of the channels generating the saccades.)

7.2 Verification results on continually updated channel weighting strategies

Compared to the previous (constant weighting) strategies, continually updated strategies assume the other extremity, namely that the triggering channels and their weights depend *solely* on the stimuli. The reality is probably somewhere in between (here the problem is, that due to the recent discovery – and modeling – of the mammalian retina channel system, there are no measurements aiming to uncover the role of the several channels *separately*, in visual attention so far – at least to the best of my knowledge).

After all, I still have expected higher efficiency from the continually updated channel weighting strategies (described in the present section) than from the constant ones. Even if the differences are not very big (according to the different approaches 1-10%) – in contrast with my expectations – the constant strategies turned out to be more effective.

The two investigated approaches are the same that have already been used in the 6th chapter: the channel weights depend on the corresponding saliency map's:
i) highest value(s)
ii) averages.

7.2.1 First approach

Following the order of chapter 6, firstly I discuss the results belonging to the approach assuming that *those channels trigger a saccade which have (one or a few) outstandingly high saliency value(s)*, anywhere in their saliency maps (see also chapters 6.1 and 6.2). Accordingly, the process has been the following:

1. Dissolve the incoming frames of the *test video set* according to the ten retina channels.

2. Create the corresponding saliency map for all the ten channels. The RFs by which these channel-dependent maps have been calculated are those depicted in table 6.2, last column.

3. Range the channels in descending order according the highest values of the corresponding saliency maps.

4. Create the master saliency map by taking into account the first i most salient channels ("saliency", in the sense of the above ranking). $i \in \{1, 2, \ldots, 10\}$. The weighting is proportional to the *maximal* values of the saliency maps. Mathematically:

$$FinalSM_k^i = \sum_{c=1}^{10} w_c SM_{k,c}$$

where

$$w_c = \begin{cases} max(SM_{k,c}) & \text{if } max\,(SM_{k,c}) \text{ is in the first } i \text{ biggest values among} \\ & \text{the ten } max\,(SM_k)\text{s} \\ 0 & \text{otherwise} \end{cases}$$

and

k	is the frame number
c	channel identifier, $c \in \{1, 2, \ldots, 10\}$. 1: Intensity, 2: Transient, 3: LED, etc.
i	the number of the channels taken into account during the calculation of the final saliency map
w_c	the weight of the cth channel (during the calculation of the final saliency map)
$SM_{k,c}$	saliency map belonging to channel c on the kth frame

5. Make *predictions*; that is, define the locations that the model marks as probable fixation locations. I have made four predictions for every frame,

that is four $(x-y)$ coordinate pairs, which were the four most salient locations of the final saliency map.

6. Compare these predictions with *measured* human fixation locations. I have defined "hit", if the distance between the predicted and the measured location was less then 5°.

Figure 7.2 depicts the results. According to the diagram, the present approach gives the best result by applying the four most salient channels (– "salient" on the given stimulus) during the calculation of the final saliency map. Further increasing of the channels number decreases the efficiency. It is also notable, that – in contrast with the expectations – the results check behind the constant channel weighting strategy with approximately 5-10% success in the case of four predictions.

The results are slightly different, if the channel based saliency maps are normalized, more accurately, if they are divided by their average. This happens between the 2nd and 3rd steps in the former detailed process. In this case, the final saliency map is defined as follows:

$$FinalSM_k^i = \sum_{c=1}^{10} w_c \cdot \overline{SM_{k,c}}$$

where the normalized map is

$$\overline{SM_{k,c}} = \frac{((IM_{k,c} * RF_r) * G_r)}{mean\,(SM_{k,c})}$$

where

$*$ denotes convolution

$IM_{k,c}$ is the activation map of channel c on frame k
(or in other words: the output of the cth channel on frame k)

RF_r r sized receptive field (r is an index)
(these are the "proper" RF sizes for each channel, summarized in table 6.2)

G_r a discrete Gauss-filter with the same size that of an r sized RF

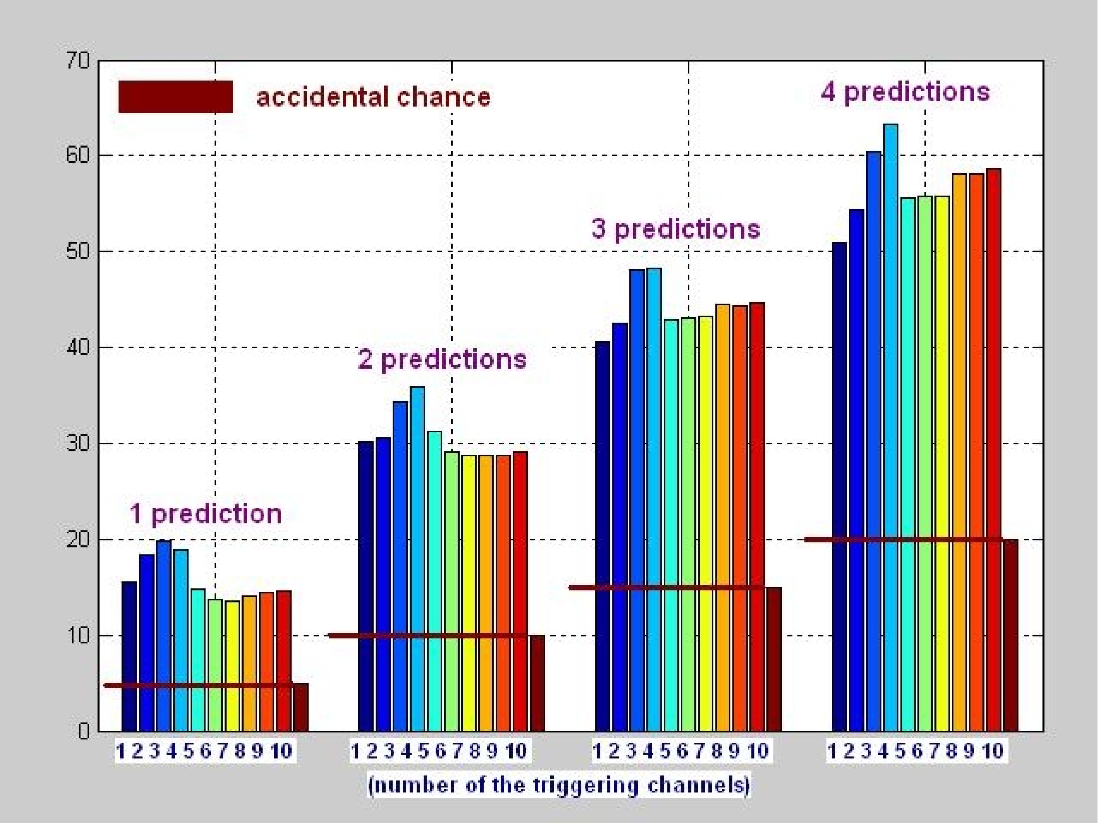

Figure 7.2: Validation results for the first approach of the continually updated channel weighting strategies, for 4 predictions/frames. Here I have assumed that a channel participates in triggering a saccade, if the corresponding saliency map contains one of the highest values that have came into existence on the given stimulus (for details see the text). The receptive fields are the ones defined previously for the corresponding approach, summarized in table 6.2. The first bar in every group shows the results for the case when only one channel creates the final saliency map (for which the highest saliency value belongs to), the second bar if two channels participate in forming the final map, etc. The horizontal bars indicate the accidental chance for making a "hit". Once the saliency maps were ready, I have made predictions referring to the locations with the highest saliency values in the master map. I defined "hit", if the distance between the predicted and the measured location was less than 5°.

The corresponding results are depicted on figure 7.3.

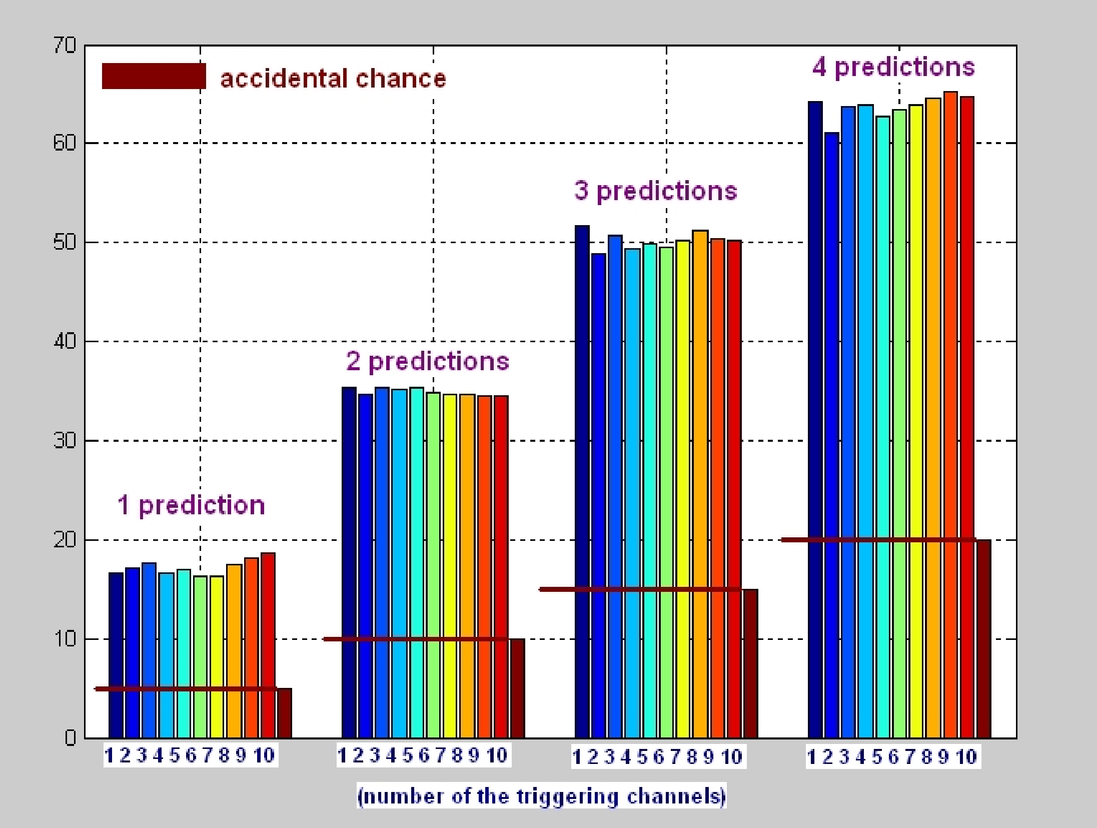

Figure 7.3: Validation results for the first approach of the continually updated channel weighting strategies, for four predictions/frames. Here the assumption has been the same one as previously (figure 7.2) *with the distinction, that the channel based saliency maps had been normalized* between the 2nd and 3rd step (see text). The first bar in every group shows the results for the case when only one channel creates the final saliency map (for which the highest saliency value belongs to), the second bar, if two channels participate in forming the final map, and so on. The horizontal bars indicate the accidental chance for making a "hit". Once the saliency maps were ready, I have made predictions referring to the locations with the highest saliency values in the master map. I defined "hit", if the distance between the predicted and the measured location was less then 5°.

As shown in the picture, in this case the accuracy is basically independent of the number of the used channels, meanwhile the accuracy is better than in the previous case, see for example the results belonging to three predictions: on figure 7.2 even the best outcome lags behind 50%, meanwhile on figure 7.3 it is around it. To conclude, the previous approach – which does not apply normalization –

mathes the present approach, if four channels' data is taken into accound during the formation of the final map.

7.2.2 Second approach

The second approach – similarly to section 6.3 – assumes that those channels contribute in the provocation of a saccade, which are salient *in average* on the actual stimuli. The algorithm alters accordingly (the altered parts are typed with small capital letters for the sake of better emphasizing):

1. Dissolve the incoming frames of the *test video set* according to the ten retina channels.

2. Create the corresponding saliency map for all the ten channels. The RFs by which these channel-dependent maps have been calculated are those depicted in table 6.3, last column.

3. Normalize the saliency maps, more accurately: divide them with their mean value:

   $$\overline{SM_{k,c}} = \frac{((IM_{k,c} * RF_r) * G_r)}{mean\,(SM_{k,c})}$$

   where the notations are the same than in section previously, defined in section 7.2.1.

4. Take out the first i saliency maps to whom the highest mean value belonged to (for all $i \in \{1, 2, \ldots, 10\}$, one after the other)

5. Create the master saliency map by taking into account the first i most salient channels in average, with proportional weighting to the average saliency values. Mathematically:

   $$FinalSM_k^i = \sum_{c=1}^{10} w_c \overline{SM_{k,c}}$$

with the same notations as previously.

6. Make *predictions*; that is, define the locations that the model marks as probable fixation locations. I have made four predictions for every frame, that is four $(x - y)$ coordinate pairs, which were the four most salient locations of the final saliency map.

7. Compare these predictions with *measured* human fixation locations. I have defined "hit", if the distance between the predicted and the measured location was less then 5°.

The corresponding results are depicted on figure 7.4.

As it can be read from the diagram, increasing the number of channels mildly increases the accuracy as well, up to the usage of 5-7 channels. The efficiency – depending of the number of the used channels – barely achieves the results of the previous approaches. For the sake of better comparison among the different continually updated strategies, figure 7.5 summarizes their accuracy. 'Mrf' denotes the strategy described firstly, belonging to figure 7.2, 'PreNormMrf' is its' altered version applying a normalization step (the corresponding figure is 7.3), whereas 'Avg' marks the latter approach described in the present chapter, to which figure 7.4 belongs to.

7.2.3 Conclusions

According to the results depicted on figures 7.1–7.5, *constant channel strategies achieved better results than the continually updated ones*, (compare figures 7.1 and 7.5). Although this is a surprise, from an engineering viewpoint it is a kind of "luck", since this approach has less computational demands.

More precisely, although the differences were not huge, the overall "winner" is the third approach from the static weighting strategies (often referred to as 'Avg'), whose accuracy abundantly exceeds 70%, on four predictions (fig. 7.1). The corresponding values are summarized in table 6.3: channel weights and receptive field sizes, all the quested parameters. The final adjustment of the described

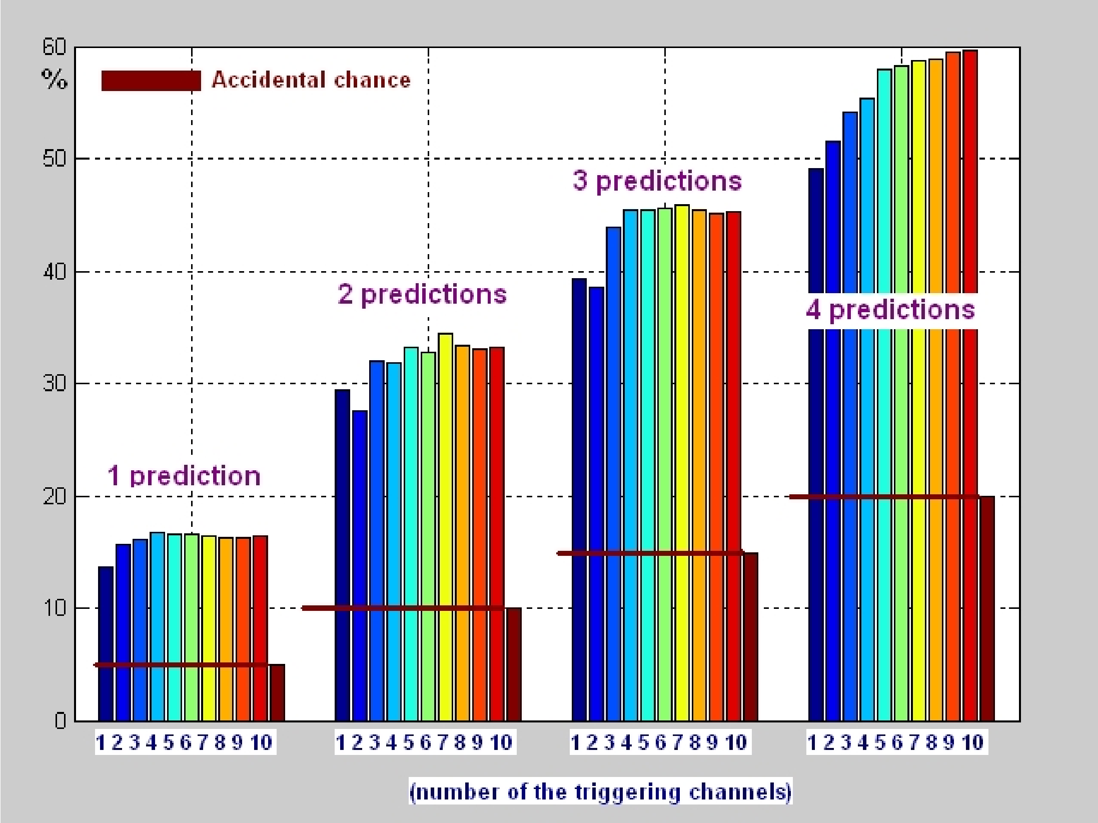

Figure 7.4: Verification data with dynamic channel choice. Here the assumption has been that a channel participates in triggering a saccade, if it is salient on the given stimulus (frame) *in average*. The first bar in every group shows the results for the case when only one channel creates the final saliency map, the second bar if two channels participate in forming the final map, etc. The horizontal bars indicate the accidental chance for making a "hit". Once the saliency maps were ready, I have made predictions referring to the locations with the highest saliency values in the master map. I defined "hit", if the distance between the predicted and the measured location was less then 5°.

attentional model has been done applying these values.

Of course – as it was already touched –, based on the validation results for the constant channel weighting strategies (figure 7.1), one might ask, why all these channel decompositions are necessary, if the Transient channel alone gives a result which is almost as accurate as the model applying all the channels? Well, as we will see in the next chapter ("Practical applications"), the role of the other

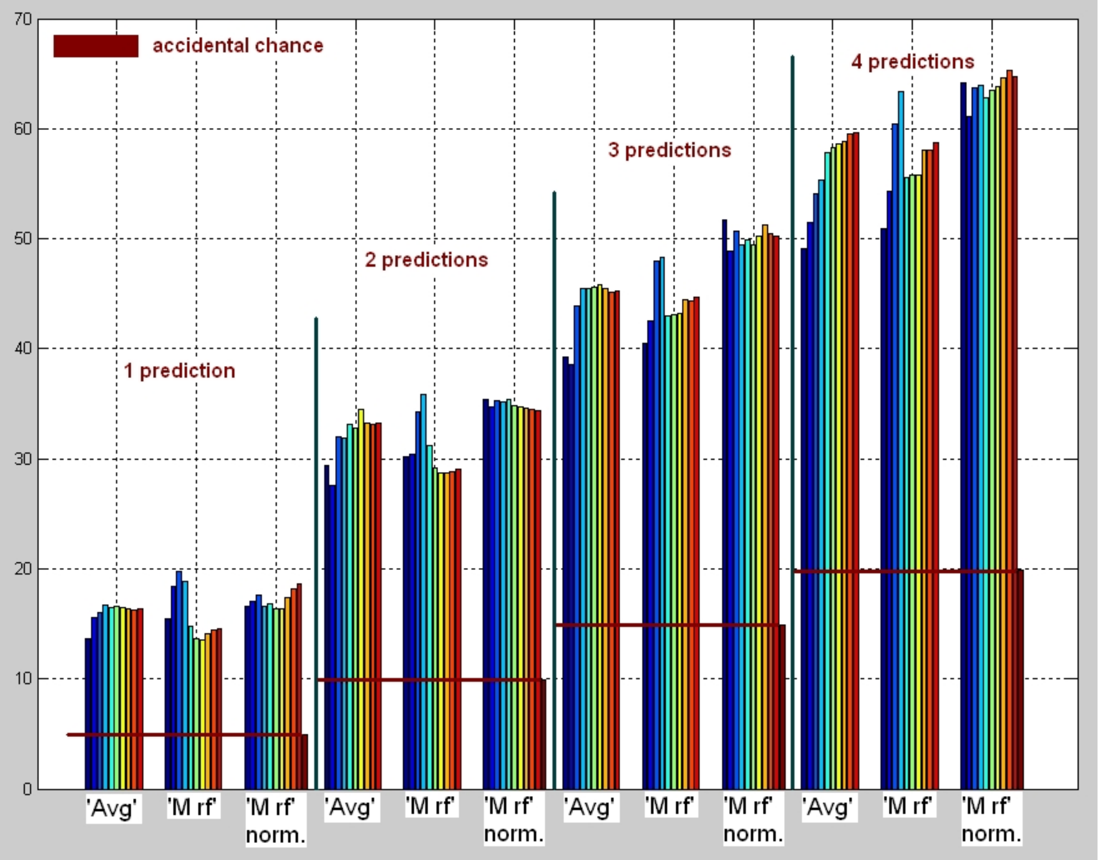

Figure 7.5: Comparison of the different dynamic channel choice strategies. 'M rf' denotes the strategy described firstly, belonging to figure 7.2, 'M rf norm' is its' altered version applying a normalization step (the corresponding figure is 7.3), whereas 'Avg' marks the latter approach described in the present chapter, to which figure 7.4 belongs to.

channels will become fundamental as soon as one tries to make any kind of further processing of the attendant data.

Chapter 8

Practical applications

Possible application fields for attentional models are extremely wide, starting from different blind navigation tasks, through robot vision up to bionic retina implants [70, 71, 72, 73]. In the present dissertation I discuss some subtasks raised in the ongoing project called “Bionic Eyeglass Project” [74, 75], which aims to help the everyday life of blind or visually impaired people. The purpose is to provide a specific kind of information, or to locate those regions in the visual scene that, with a high probability, contain important information for visually impaired people - that is: to define the Region of Interest (“ROI”) in an unstable, low resolution video input, recorded by the visually impaired person.

In this stage of the project, the input comes from a mobile phone’s video camera in 176x144 pixel resolution (called Q-CIF), but the phone is now being extended by a Cellular Visual Microprocessor (which is the Q-Eye in the Eye-RIS system, a product of the AnaFocus Ltd., Seville, or the Bi-i camera computer of Analogic Computers Ltd., Budapest). The diversity of interesting tasks as well as the construction of the required database has been compiled with the help of members of the “Hungarian National Association of Blind and Visually Impaired People” [74]. In the thesis I present efficient new algorithms for:

- Finding light sources (lamps) – this task (although it seems to be a trivial ‘problem’ for a person with normal vision), could prevent lots of annoyance for visually impaired people, for example, by preventing the lamps to remain switched-on for weeks after a guest.

Here, the most important criterion is that the solution has to be independent from the input's actual brightness, that is, the accuracy should be the same in the case of a sun-drenched and a dark room.

- Locating LED indicators (in real-life indoor and outdoor scenes).
- Finding traffic signs in real-life street scenes.
  The main purpose of these two latter tasks is to realize a fast method that locates the areas which contain the traffic signs / LED indicators with high probability, on complex real-life outdoor scenes. Subsequently, a classifier algorithm has to analyze only the located ROIs instead of the whole input, which can fasten up the whole process significantly. The main difficulties derive from the instability of the by-default bad-resolution input, the unconstrained lighting conditions, and from the variety of the possible inputs.

The algorithm's main functional components are: video stabilization, retina channel decomposition (or "low-level feature extraction" – see section 3.2 and Appendix A), and saliency map generation. A summarizing flow chart can be seen on figure 8.1, which has been taken from the publication [76], in which I report of the present algorithm.
The input of the *whole* process is an image flow taken by a mobile phone extended with a Cellular Visual Microprocessor, and the output consists of audio information for the person using the equipment. In the present thesis I do not deal with the methodology of transformation of the demanded information into audio format, but with the problem of locating the demanded information within a video flow.

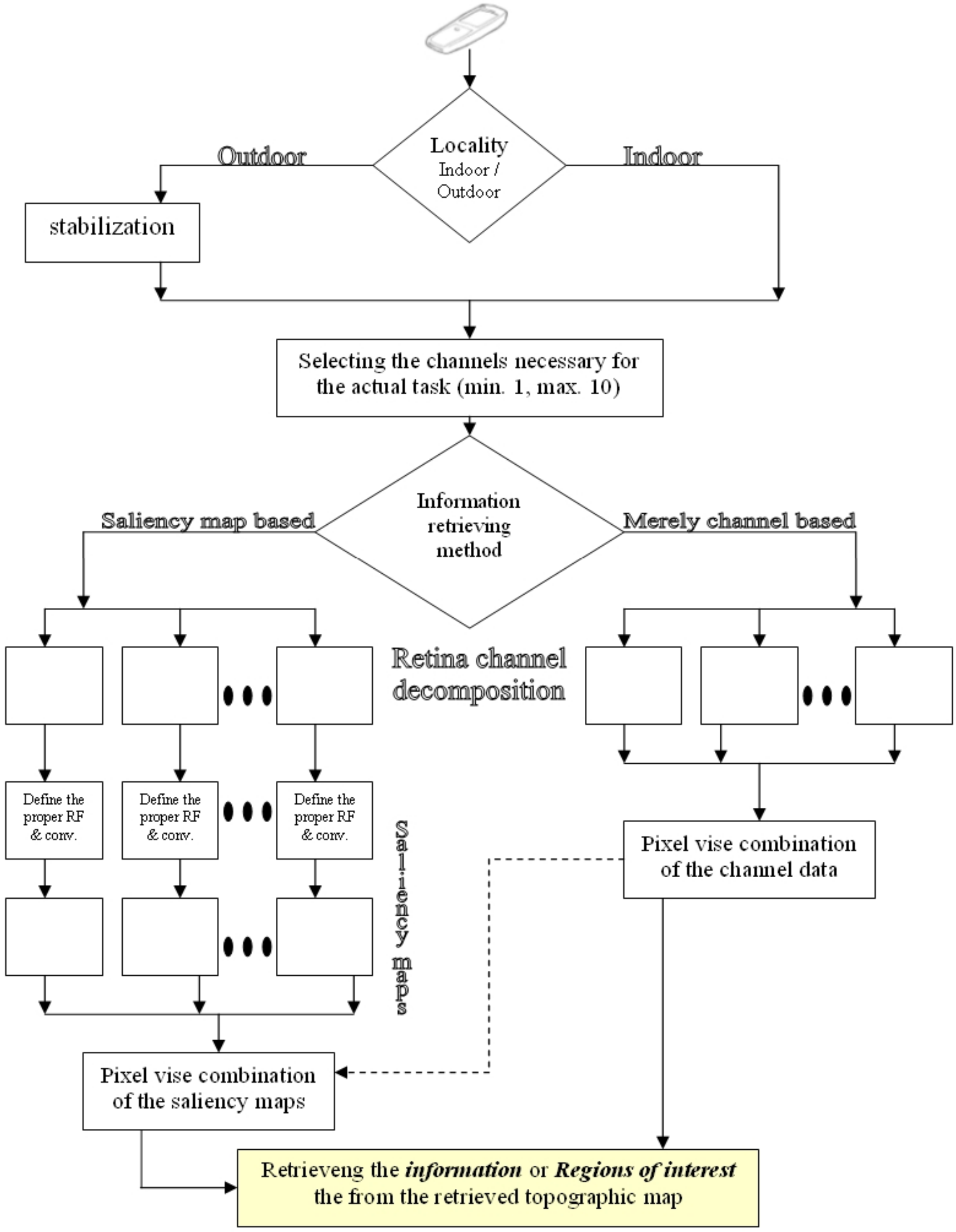
Locality
Indoor / Outdoor
Outdoor
Indoor
stabilization
Selecting the channels necessary for the actual task (min. 1, max. 10)
Information retrieving method
Saliency map based
Merely channel based
Retina channel decomposition
Define the proper RF & conv.
Define the proper RF & conv.
Define the proper RF & conv.
Saliency maps
Pixel vise combination of the channel data
Pixel vise combination of the saliency maps
Retrieveng the *information* or *Regions of interest* the from the retrieved topographic map

Figure 8.1: The flow chart of the proposed method. The input is a strongly unstable, low resolution video flow coming from a mobile phone's camera held by a visually impaired person. The output can be: **Regions of Interest** (e.g. locations of LED indicators, traffic signs), or **Specific information** (e.g. if there is any switched-on lamp or not). The dashed line shows an optional information combination step (raised in the task of locating traffic signs, where retina channel data and saliency map data had been combined.)

8.1 Stabilizing the input

Image flows provided by a camera held by a blind walking person are usually extremely noisy and unstable, often accompanied by fast, unexpected camera motions. The recording equipment (camera) can be

- rotated
- shifted in the vertical and horizontal direction, and
- transported in the direction of motion (e.g. looming).

Additionally, often the picture's main objects shift significantly from one frame to another, e.g. during turning around. The goal of the image stabilization step is to keep the steady objects (e.g. buildings) in the same pixel positions, while the moving objects (for example the pedestrians) can change position. (In-built standard camera image stabilizers – both mechanical or digital solutions – can only handle much smaller dislocations.)

It is useful to define the transformation-parameters between *adjacent* frames, instead of estimating the difference between the reference frame and the actual frame. In this manner, it is possible to trace bigger deformations throughout longer frame-series. Then, the calculated transformation 'inherits' from frame to frame, as follows:

If the actual reference frame is the ith one, then the $P_{i+k,i}$ vector contains the transformation parameters between the actual frame $i+k$ and the reference frame i. In the next step, the vector $P_{i+k+1,i+k}$ is calculated, which contains the transformation values between the actual adjacent frames: $i+k+1$ and $i+k$. Then $P_{i+k+1,i} = P_{i+k,i} \oplus P_{i+k+1,i+k}$ will be updated, and will comprise of the differences accumulated throughout the $k+1$ frames that have been captured since the last reference-frame updating.

Figure 8.2 depicts the flow chart of the stabilization. (See also [77, 78]) The key element in it, is how the transformation parameters are defined (I have highlighted this step with a bit darker shade on the diagram). The following sections

(8.1.1 and 8.1.2) targets this question.

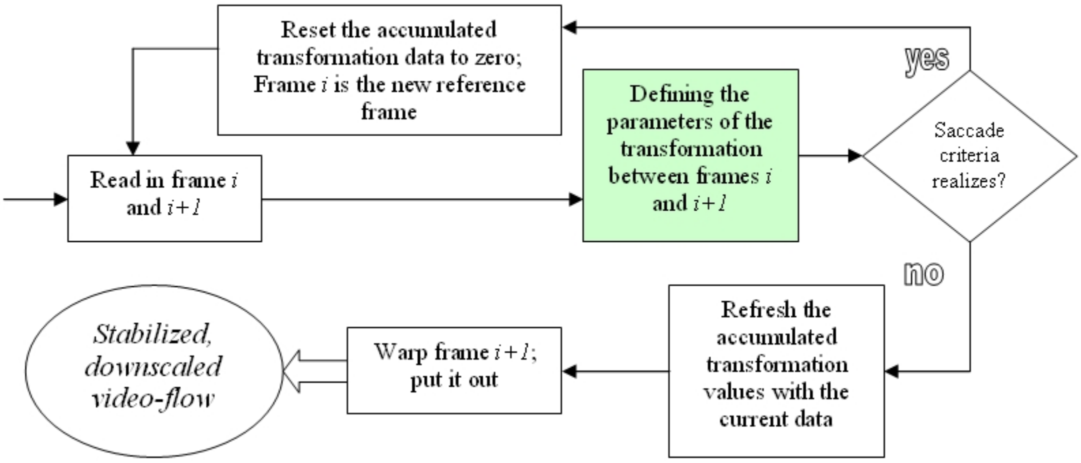
Reset the accumulated transformation data to zero; Frame *i* is the new reference frame
Read in frame *i* and *i+1*
Defining the parameters of the transformation between frames *i* and *i+1*
Saccade criteria realizes?
yes
no
Refresh the accumulated transformation values with the current data
Warp frame *i+1*; put it out
Stabilized, downscaled video-flow

Figure 8.2: The flow chart diagram of the stabilization. The input (left hand side, top of the picture) is an unstable video-flow coming from a mobile phone's camera. The output of this algorithm is the stabilized video flow (left hand side, bottom of the picture. The goal is to keep the steady objects (e.g. buildings) in the same pixel positions, while the moving objects (e.g. pedestrians) can change position.

8.1.1 Mathematical background

To estimate the instantaneous velocity field I have modeled the motion image by a continuous variation of image intensity as a function of position and time. The intensity value on position (x, y) at time t is described by the $f(x, y, t)$ intensity function.

If we expand this function in a Taylor series, we get:

$$f(x + dx, y + dy, t + dt) = f(x, y, t) + \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial t}dt + HT \qquad (8.1)$$

where HT is for higher-order terms, which are usually ignored [79].

The crucial observation that is exploited, is that if the image at some time $t + dt$ is a result of the original image at time t being moved translationally by dx and dy, then

$$f(x + dx, y + dy, t + dt) = f(x, y, t). \qquad (8.2)$$

Thus, from equations (8.1) and (8.2) we get:

$$0 = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial t}dt \tag{8.3}$$

or, in other form:

$$-\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} \tag{8.4}$$

$\frac{\partial f}{\partial t}$, $\frac{\partial f}{\partial x}$, and $\frac{\partial f}{\partial y}$ are measurable quantities, while $\frac{dx}{dt}$ and $\frac{dy}{dt}$ are the quested values, namely the velocity in x and y directions.
Using the $\frac{dx}{dt} = u$ and $\frac{dy}{dt} = v$ notation, we get

$$-\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}u + \frac{\partial f}{\partial y}v \tag{8.5}$$

or, equivalently,

$$-\frac{\partial f}{\partial t} = \nabla f \cdot \mathbf{u} \tag{8.6}$$

where ∇f is the spatial gradient of the image and $\mathbf{u} = (u, v)$ is the velocity vector.

8.1.2 The detailed course of the calculation

–*Measured values:* $I_x(= \frac{\partial f}{\partial x})$, $I_y(= \frac{\partial f}{\partial y})$, and $I_t(= \frac{\partial f}{\partial t})$, the intensity gradients
–*Calculated values* (the estimations): $u(= \frac{dx}{dt})$ and $v(= \frac{dy}{dt})$.
With these notations (8.5) will be, for every pixel:

$$I_x u + I_y v + I_t = 0 \tag{8.7}$$

.

The measured values

The I_x, I_y *spatial gradients* can be determined by convolution, where the kernel is the $\begin{bmatrix}-1 & 8 & 0 & -8 & 1\end{bmatrix}/12$ vector, which is a commonly used estimation in the literature [80]. (This kernel is applied on the Gauss-filtered image.)
The I_t *time gradient* is simply the difference between the two Gauss-filtered images.
(As it follows from the above process, these values are defined for each and every pixels, so I_x is not a scalar, but a matrix, and I_y, I_t similarly.)

Defining the parameters of the projection:

For *mapping function*, I have chosen a linear affine transformation, which can handle shifts in x and y directions, scaling, rotations and shears, as follows:

$$u = a_0 + a_1 x + a_2 y \tag{8.8}$$

$$v = b_0 + b_1 x + b_2 y \tag{8.9}$$

where a_0, a_1, a_2, b_0, b_1 and b_2 are the parameters of the transformation, which we want to determine. In the matrix-form of (8.7), the meaning of these parameters can be understood better:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} + \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{8.10}$$

In this equation $\begin{bmatrix} a_0 \\ b_0 \end{bmatrix}$ defines translation in x and y directions, while $\begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$ describes the scaling, rotation and shear.

Thus, from (8.7), (8.8) and (8.9), for every pixel we get:

$$(a_0 + a_1 x + a_2 y) I_x + (b_0 + b_1 x + b_2 y) I_y + I_t = 0 \tag{8.11}$$

which will be

$$\begin{bmatrix} I_x^{(1)} & I_x x^{(1)} & I_x y^{(1)} & I_y^{(1)} & I_y x^{(1)} & I_y y^{(1)} \\ I_x^{(2)} & I_x x^{(2)} & I_x y^{(2)} & I_y^{(2)} & I_y x^{(2)} & I_y y^{(2)} \\ \ldots \\ I_x^{(k)} & I_x x^{(k)} & I_x y^{(k)} & I_y^{(k)} & I_y x^{(k)} & I_y y^{(k)} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ b_0 \\ b_1 \\ b_2 \end{bmatrix} = - \begin{bmatrix} I_t^{(1)} \\ I_t^{(2)} \\ \ldots \\ I_t^{(k)} \end{bmatrix} \tag{8.12}$$

where k (the number of rows) is the number of pixels.

With the notations:

$$A = \begin{bmatrix} I_x^{(1)} & I_x x^{(1)} & I_x y^{(1)} & I_y^{(1)} & I_y x^{(1)} & I_y y^{(1)} \\ I_x^{(2)} & I_x x^{(2)} & I_x y^{(2)} & I_y^{(2)} & I_y x^{(2)} & I_y y^{(2)} \\ \ldots \\ I_x^{(k)} & I_x x^{(k)} & I_x y^{(k)} & I_y^{(k)} & I_y x^{(k)} & I_y y^{(k)} \end{bmatrix} \tag{8.13}$$

and

$$P = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ b_0 \\ b_1 \\ b_2 \end{bmatrix} \tag{8.14}$$

we get

$$A \cdot P = -I_t \tag{8.15}$$

from where

$$P = A^{+} \cdot (-I_t) \tag{8.16}$$

where '+' denotes pseudo-inverse, and the P vector contains the transformation-parameters, which we were looking for.

8.2 Retina channels, saliency maps and receptive fields in the applications

Practically, from an engineering viewpoint, a saliency map is a retina channel output (or the result of the 'low level visual feature extraction') convolved with a receptive field. RFs can be represented in matrix form.
Figure 8.3 depicts how I have determined the optimal RF in the task of finding traffic signs, as an example of their usage.

- The *size* of the receptive field is determined as follows:
  - from one hand, the viewing-angle of the mobile phone is ˜45°, which occupies 176 pixels. This means, that roughly 3.9 pixels cover 1°.
  - on the other hand, at the video flow, the size of an object depends on its' distance. Namely, its' size in viewing angle is $\tan\frac{\alpha}{2} = \frac{ObjectRadius}{distance}$ (figure 8.3 a). Thus, the sensing-distance of a 45 cm diameter (0.225 m radius) traffic sign from ˜7-8 m covers a bit more than 3.3°. Thus, the optimal inner diameter of the receptive field is 13 pixels. (figure 8.3 ***a*** and ***b***)

- The outer size of the RF has been adjusted according to 'real' RFs in which the inner part covers around the half of the whole RF in viewing angle.

- The values of the matrix have to satisfy the criteria of
  - giving maximal response if antagonistic stimuli hit the RF's inner and outer area
  - giving no answer if the input image-region contains equal values – that is, in case of uniform lighting

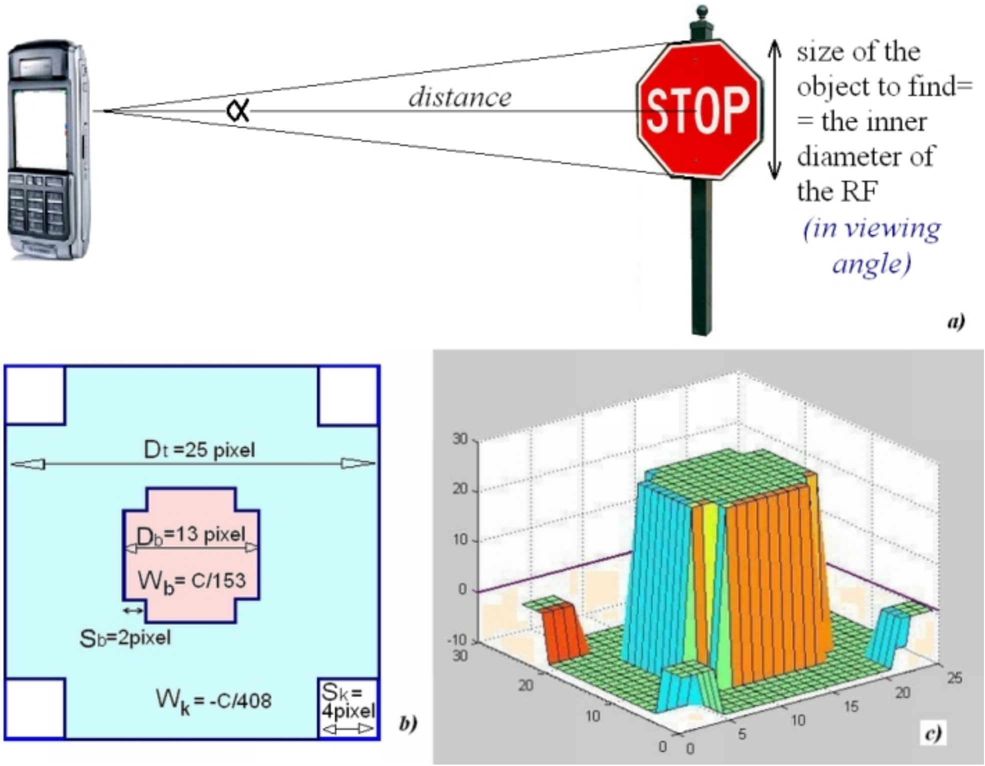

Figure 8.3: Receptive field adjusted for the task of finding traffic signs. Figure ***a***: the inner diameter of the receptive field and the size of the searched object should be the same in viewing angle. This criterion helps to determine the size of the RF. ***b*** and ***c***: the resultant; On ***c*** the height and the depth are proportional to the weights, which have – due to the antagonistic behavior of the RF's inner and outer part – opposite sign. The zero level is emphasized with purple line.

The maximal value is arbitrary, since it is only a constant multiplier ("C" on figure 8.3 ***b***).
In the followings I get on the subject of carrying out the specific tasks, based on the foregoing.

8.3 The Realized Applications

8.3.1 Locating traffic signs

The main purpose of the present algorithm is to realize a fast method locating the areas which contain traffic signs with high probability, on complex real-life outdoor scenes. The main difficulties derive from the instability of the input – which has by default bad resolution – and from the fact that the lighting circumstances can vary on a wide range. Traffic signs – due to their color and shape design – can effectively be detected by circle-shaped receptive fields on color opposition channels. For solving this task, I have used the RF determined in the previous section (figure 8.3) on the Blue-Yellow color opposition channel. According to the experiments, rooftops and building walls often effectuate salient areas with the blue sky, something that can lead to false results. In order to avoid these errors, I have applied the Delta channel's data as well: since it gives a vivid response on light sources, only those regions have been taken into account, where this channel has given a smaller response than a given threshold.

Figure 8.4 shows some typical frames from the test database. The distortions of the input frames are due to the stabilization method. Important to note, that this method does not exploit any additional information or knowledge (for example, that traffic signs are primarily expected in a given height), thus with the guidance of the equipment the results can be further improved. According to the test results, the main error sources have been: from the one hand, shadow, which leads to false negative results because of the loss of color information (figure 8.4 c), and from the other hand, objects with 'appropriate size' and vivid colors, which lead to false positive results (figure 8.3 d). Important to note, that – because of the lack of a commonly accepted test database for these problems –

the evaluated information significantly depends on the test database. Tables 8.1 and 8.2 show the test results.

(Frame percentage)	Correct answers	False answers
There **is** traffic sign on the input video frame (total 502 frames)	**73.7%** (370 frames out of 502)	**26.3%** (132 frames out of 502)
There is **no** traffic sign on the input video frame (total 414 frames)	**95.4%** (395 frames out of 414)	**4.6%** (19 frames out of 414)
Total (916 frames)	**83.5%** (716 frames out of 916)	**16.5%** (151 frames out of 916)

Table 8.1: The results for the task: "Locating traffic signs". "Correct" answer means that *either* the input frame has no traffic signs on it and there are no located areas on the output either, *or*, there is at least one sign on the input and there are located areas on the output as well. The test video set included 916 real-life frames from different locations and with different lighting conditions.

Since Table 8.1 does not indicate the *accuracy* of the located areas ("ROIs"), I provide another table (8.2) showing these results.

	Correctly identified locations	Incorrectly identified locations
Altogether 490 located areas	**73.7%** (361 ROIs out of 490)	**26.3%** (129 ROIs out of 490)

Table 8.2: The accuracy of the identified locations. Only those frames are included, where there was at least one located area. A ROI is "correct" if there is a traffic sign at that very location, and "incorrect" otherwise. Thus, an answer belonging to one single frame can contain both correct and incorrect locations (see for example figure 8.4 d).

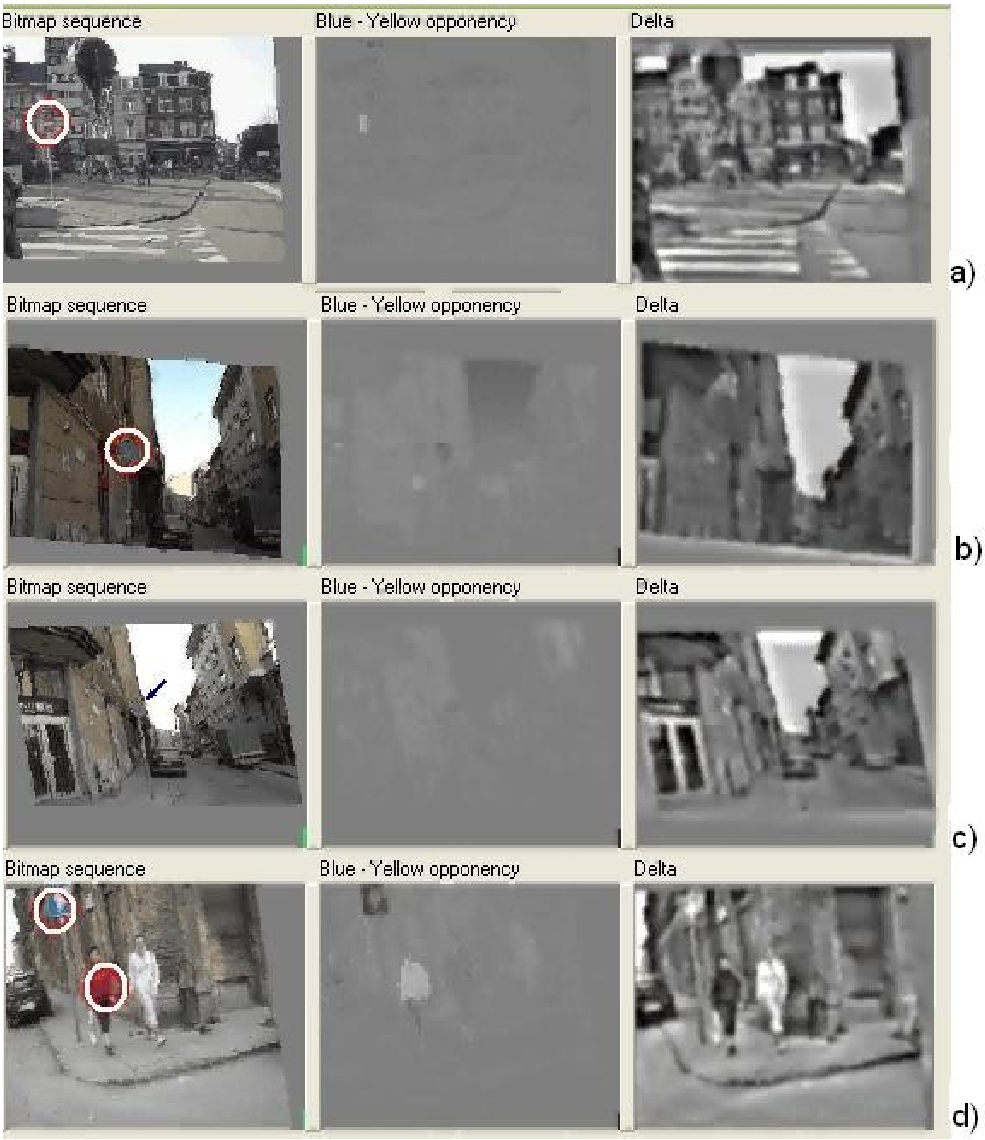

Figure 8.4: Some typical frames from the test database I have used to evaluate the task aiming to locate traffic signs. In all the four rows, the left-most image is a frame from the input video flow with the areas identified as traffic signs (white circles). The other two images in each row are the corresponding outputs of the used channels: the middle ones are the Blue-Yellow color opposition channels' output and the right ones are the response of the Delta channel. ***a*** and ***b*** are examples for correct results. ***c***: The prime cause of the false negative answers (when the sign is not located) was due to the loss of the color information, which happened when the sign was in shadow. On figure ***c*** the blue arrow points to a traffic sign being in shadow. (These signs are difficult to see even with "pure eyes") From closer they can be identified: ***b*** is the same as ***c*** from a few meters nearer. ***d*** depicts the typical reason for false positive results: vivid colors with the "appropriate size". The input frames are distorted because of the stabilization. Table I and II indicate the test results.

8.3.2 Finding lights sources

A trivial matter for people with normal vision but often a hard task for the blind ones, is to detect whether the lamps are switched on or switched off – for example after leaving guests. According to our consultant from the 'Hungarian National Association of Blind and Visually Impaired People' [81] – with whom the tasks has been defined together – an algorithm solving this task could prevent much annoyance.

Here, the most important criterion is that the solution has to be independent of the input's actual brightness, that is, its' reliability should be the same in the case of a sun-drenched room and a dark cell.

The solution for this subtask differs from the former one in the sense that here I rely merely on retina channel information – instead of saliency maps. One channel proved to be enough for this task, namely the "Polar" channel, which seems to respond on light sources [9, 21] (figure 8.5). It gives strong reaction on primary light sources, both for natural (sun) and artificial ones (lamps) – and, to reflecting surfaces as well (mirrors, glass-tables, etc.) which cause a small error rate. Still, the accuracy this channel enables is very high: the ratio of the correct answers reaches 98-99% (see table 8.3).

Since this channel responds to natural light sources as well (figure 8.3 e), the user is supposed to know where the window is, but this is not a real restriction in every-day practice, since people usually do know the location of the windows. Otherwise, precise knowledge about the location of the lamp(s) is not a demand, since the visual environment can be scanned.

According to the experiments, the Polar channel saturates (gives maximal response) on those areas where primary light sources are present, and give no answer elsewhere (figure 8.5). It follows that the accuracy of the algorithm is completely independent of the quality of the input video-flow, and also, it does not depend on the *brightness* either. The results depicted in table 8.3 are based on test videos made in sunshiny rooms.

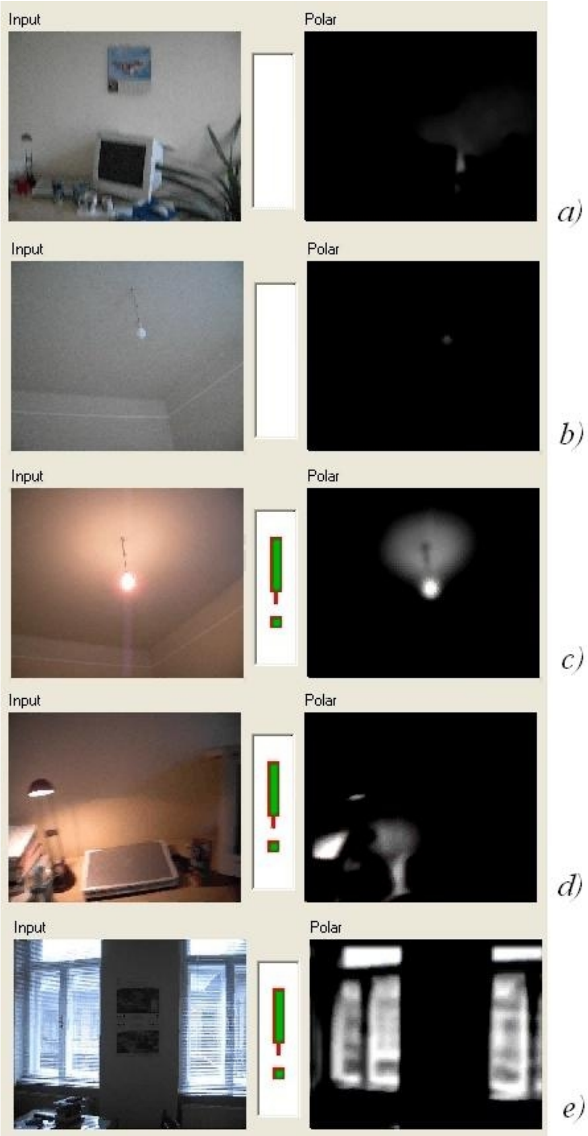

Figure 8.5: The “Polar” channel responds on light sources, both for natural (***e***) and for artificial ones (***c, d***). The pictures are taken from the test video set. All the five figures show the input on the left, and the corresponding output of the Polar channel, on the right. ***a*** and ***b***: a part of a bright room in day light; the Polar channel is basically silent. ***c*** had been recorded a few seconds after ***b***: the lamp is switched on, the Polar channel is excited. (The exclamation mark between the two channels indicates that the answer is: “There is light source on the input!”) The Polar channel enables very high reliability for this task (see table 8.3).

	Correct answers	False answers
There **is** light source on the input video frame	98.8%	1.2%
There is **no** light source on the input video frame	99.38%	0.62%

Table 8.3: The test results of the algorithm aiming to detect primary light sources, independently from the brightness of the input, or in other words, from the intensity values. The process is based on one of the mammalian retina channels (namely the "Polar" channel), which reacts on light sources. The small error is due to reflecting surfaces (a glass table in our case). These values are based on the evaluation of test videos made on shiny rooms, including 1563 frames together.

8.3.3 Locating LED indicators

In many public buildings, offices and transport vehicles basic information is transmitted by LED indicators. The aim of this method again is to carry out a fast solution that localizes the areas that contain the indicators in question with high probability, on various indoor and outdoor real-life scenes. The main difficulty – over the bad resolution and the instability – originates from the variety of the possible inputs.
The test video set I have used includes multifarious scenes including different public and private places. Some of these can be seen on figure 8.6 and 8.7. On figure 8.6 I have also visualized those three channels that I have used for solving this task. These are the two color-opposition channels (blue-yellow and red-green, which are used because LED indicators are colored) and the Delta channel, which – according to the experiments – proved to be the most appropriate for the given task. The function of this channel has not yet been precisely formulated up to present, but according to the observations, it gives significant response for small or fragile light sources as well (similarly to strong light sources).
From here, the selection algorithm is the following: if on a given location at least one of the two color opposition channels gave bigger response than a certain threshold, then a "fitness-value" would be calculated, being directly proportional to the three channel-data at the given point. Afterwards these values would be arranged into descending order, and the first few locations would be the solution for the given frame, that is regions that the algorithm defines as presumptive LED locations.

Table 8.4 shows the results I have measured on this task. The results are based on the evaluation of 1207 frames. I have tested the method on real-life scenes, taken from different areas with various lighting conditions, reflecting areas, light sources, colors, etc. As it turned out, the algorithm is not sensitive to the quality of the input (e.g. resolution), to the lighting conditions or colors, either to the reflecting areas, but it is sensitive to colored lamps (see figure 8.7 b) – which is not surprising since LEDs basically are small colored lamps, until no further object or pattern recognition algorithm is used.

(Frame percentage)	Correct answers	False answers
There **is** LED indicator on the input video frame (total 951 frames)	**96.6%** (919 frames out of 951)	**3.4%** (32 frames out of 951)
There is **no** LED indicator on the input video frame (total 256 frames)	**41%** (105 frames out of 256)	**59%** (151 frames out of 256)
Total (1207 frames)	**84.83%** (1024 frames out of 1207)	**15.17%** (183 frames out of 1207)

Table 8.4: The results for the task: "finding LED indicators". The values are based on the evaluation of 1207 frames. First row first column is the correct positive (96,6%), second row first column is the correct negative result (41%). The test database included complex real-life scenes with different lighting conditions, colored and reflecting areas and colored lamps. As it turned out, the algorithm in *not* sensitive to the quality of the input (resolution), to the lighting conditions and colors, either to the reflecting areas, but it *is* sensitive to colored lamps – which the few frames (total 256) that did not contain LED happened to teemed in. The bad results are due to these lamps (figure 8.7 b). In the third row ("Total"), all the frames are counted, that is, "correct answer" indicates the percentage of the frames where either the input included LED indicator (one or more) and the output was at least one located area, or the input did not include LED indicator and the output had no located areas. Accordingly, the line "False" indicates the rest. The percentage means *frame* percentage.

Since the input frame may contain more than one LED indicator, and also, the output can be more than one located region (see figure 8.6 and 8.7), the evaluation – similarly to the task of finding traffic signs – is not as straightforward as in the previous task, where the answer was binary ("there IS light source on the input"/"there is NO light source on the input"). Thus I give another table as well, which indicates the *correctness* of the locations which the algorithm has given as solutions. In contrast with table 8.4, table 8.5 depicts *ROI percentage* instead of *frame percentage*, that is, the ratio of the correct and false located areas. Only those frames are included, where there was at least one located area.

	Correctly identified locations	Incorrectly identified locations
Altogether 2075 located areas	**81.36%** (1688 ROIs out of 2075)	**18.64%** (387 ROIs out of 2075)

Table 8.5: The accuracy of the identified locations. Only those frames are included, where there was at least one located area. A ROI is "correct" if there were a LED indicator at that very location, and "incorrect" otherwise. Thus, an answer belonging to one single frame can contain both correct and incorrect locations (see for example figure 8.7 b, where the marking of the colored lamp is incorrect (left hand side, top of the picture), while the sign on the elevator panel is correct (right hand side, top of the picture).

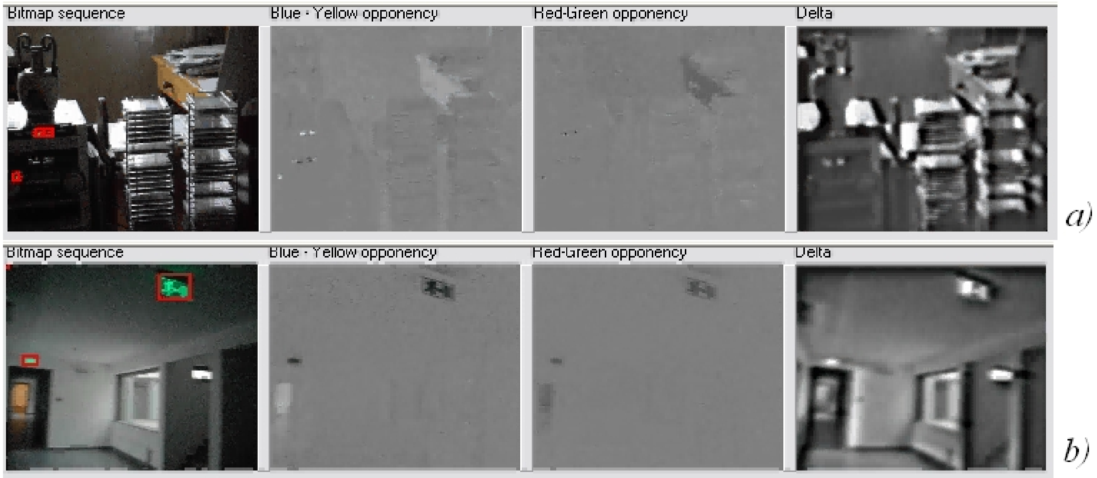

Figure 8.6: Two frames of the test database for the task "finding LED indicators". The left-most pictures in both lines show the input with the identified locations on them. The other three pictures belong to those channels, whose data has been used in the execution of the task. These are the red-green and blue-yellow color opposition channels and the Delta channel. (***a***) LEDs belonging to a hi-fi set in a room. The various reflecting surfaces do not confuse the algorithm. (***b***) corridors in the university.

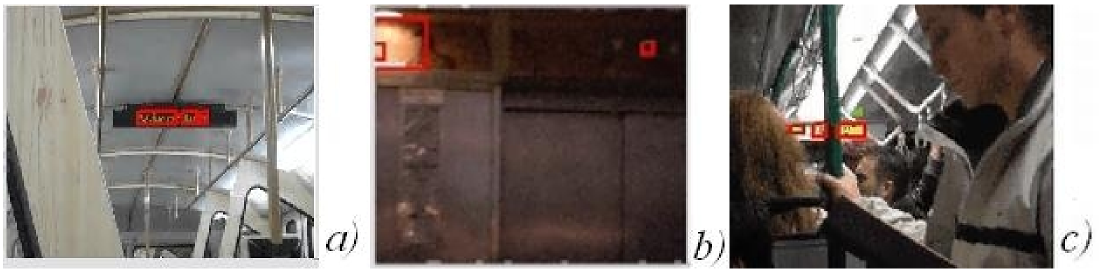

Figure 8.7: Some frames from the videos that I have used for testing the algorithm that finds LED indicators. (***a***) rack-railway from the inside (these indicators show the name of the next stop and the actual time) (***b***) a colored decorating lamp (left; the typical error source) and a LED indicator (right) showing the floor-number on a lift-panel in a department store. (***c***) tram interior.

8.4 Corresponding future tasks

In the followings I briefly mention some possible directions referring to how the introduced applications could be improved. These ideas have risen during the realization and evaluation of the different tasks.

- During walking, a camera held in a hand, makes a quasi-periodic motion. Most of the people have their own way of “swinging” the phone, thus the transformation-parameters (vertical/horizontal shifts, the angle of the rotation, etc.) characterize the certain users. These quasi-periodic parameter values could be learned during a certain amount of frames (and could even be adjusted during the entire usage), thus they become predictable for a given user. In this manner, by taking the predicted transformation values into account, the quality of the stabilization can be improved.

- The model described in this chapter is attentional in the sense that it locates regions on the input where something important appears. Naturally arises the possibility of applying a more elaborated pattern or object recognition algorithm onto the selected area.

- Many possibilities lie in the retina channel decomposition. Thus, the further investigation of the individual channels can lead to a promising basis for different scene analyzer and object recognition algorithms. For example, some time-dependent channel (primarily the Transient, Beta and the Bistratified channels) seem to play an important role in separating the different objects from each other – although, this area needs further investigations.

- In a more elaborated version, the threshold values can be adjusted by a learning algorithm, and also could be adaptive according to the different scenes and tasks.

8.5 General Conclusions and Perspectives

Visual attention – both the neurobiological background and the corresponding engineering models – is a very actively researched area, in which new results are reported basically every day. The main reason behind this is that new equipment used in biological research, (such as EEG, MEG, MRI, fMRI, PET, etc. [82]) for the first time in the history of science allows specialists to reveal more and more details of the underlying biological structures of this most important sensor of ours.

The interest from the engineering viewpoint is obvious: the range of the possible applications is extremely wide: from robot vision through different defense and observing equipment up to bionic implants, the variety is huge. Moreover, there is a very stimulating feedback between biological research and the engineering principles and applications, as the one interacts with the other via new ideas and directions.

Regarding the visual attentional models, the realization of a *complete, neuromorphic* one, is something held by the far future, since the neurobiological understanding of the top-down method is in its very infancy – just to mention some of the relating categories: scene *understanding*, object recognition, visual *consciousness*, etc.

Until then – although, according to the experiences, the direction of science is widely unpredictable – different, more or less bio-inspired or neuromorphic task-specific algorithms will be developed. The ones including TD functions will be necessarily *heuristic*.

The rough flow-chart of a possible model is given hereinafter, which can even be interpreted as an 'extension' of the Bionic Eyeglass. It operates in the same three main modes, as human TD attention can: object based, spatial based and feature based [24]. The user can switch among these modes:

- Spatial based: process all the features that are present at a given location of the visual scene, and try to recognize it. Regarding the question: "What is there?"

- Feature based: finding locations in the entire visual scene which contain a given feature or feature-collection. Like "Find locations containing 'brown' AND 'oblong' objects." Meaning: Where is my suit-case? Generally speaking, regarding the questions: "Where is the... ?", if the features of the searched object are known, that is, it is in the 'Database of the known objects'.

- Object-based: process (and store) the features of a given object (for example with the purpose of memorizing it, that is, saving it to the 'Database of the known objects'.)

This model gives two outputs: firstly, the small region of the visual scene that is worth further investigation (that is, focus of attention), and secondly, the recognized objects.
Of course, the efficiency and practicability of the above model lies in the details how the higher-order tasks (like "object investigation", etc.) are realized – here, simply represented by boxes.

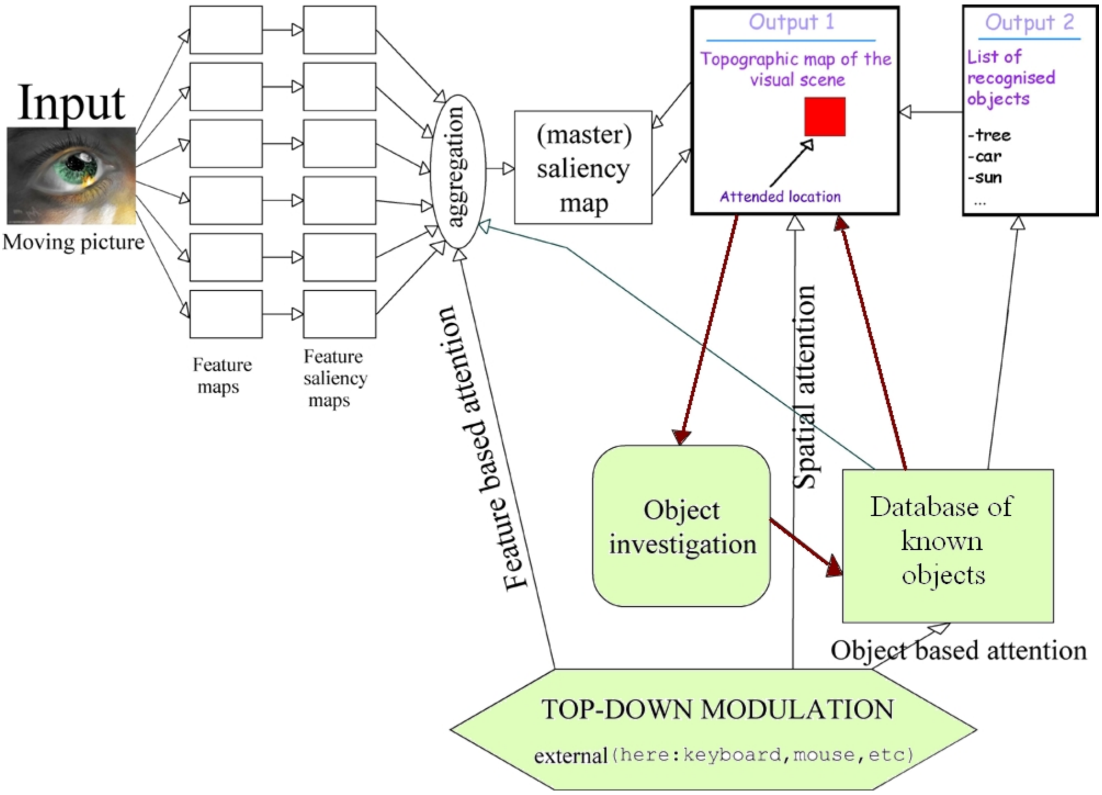

Figure 8.8: A rough flow-chart of a possible example of a heuristic, complete visual attentional model. ("complete" in the sense, that enables functions that are bounded to the top-down method). It can operate in the same three main modes, as human TD attention can: object based, spatial based and feature based. The user can switch among these modes. The 'pure' bottom-up part is in the top, left-hand side of the figure. The arrows composing the important loop among the focus of attention, the function of object investigation, and the database of the known objects, are shown bold. The known objects are identified via their features. If the features do not fit exactly, further investigation (data collections) might be necessary from the surrounding visual scene. If the features do fit with a high accuracy, the object is identified.

Chapter 9

Summary

9.1 Experimental methods

My research area requires the joint application of different disciplines. Accordingly, as a first step, via neurobiological studies I have got acquainted with the basics of vision and with the mechanisms that form visual attention as well.

The substance of *modeling* lies in the proper selection of the elements forming a complex system (like an animals' visual system), more precisely, the *selection of those elements* which develop the features being important for us. Thus, if these elements are the same in different systems, then trespassing is possible among these systems. Accordingly, in a general sense, the vertebrate visual system can be considered as the basics of my model. (Present-day attentional models are by far not precise enough to detect the differences, for example between humans and primates.)

The main steps of the model are depicted on figure 4.1. As a first step, the input image (left hand-side, top) is being decomposed according to ten different retina channels (right hand-side, top). Next, each channel creates its' own saliency map, which is a two dimensional topographic map of the physical world in the brain (right hand-side, bottom). The weighted sum of these maps form the "final" or "master" saliency map (left hand side, bottom), which is a topographic map of the visual scene as well. The saliency map codes how striking, how obtrusive are the corresponding points in the physical world. The most intense point of this map attracts our attention the most, thus the corresponding location of the physical world is being mapped into the center of the sharp seeing, that is, to the fovea.

I have realized the above model in (Borland) C++.
The first main step has been the investigation and the completion of the retina channels. The model runs on a CNN (Cellular Neural/Nonlinear Network) simulator, which I have also prepared in Borland C++. I have got the proper parameters, which define the exact spatio-temporal behavior of the different retina channels, from a previous work carried out by David Balya. Further developments of the model – of course, under the guidance of my supervisor and consultant – constitute my own work.

The principles underlying the retina-model are briefly the following: every retinal cell-layer (photo-receptors, horizontal, bipolar, amacrine and ganglion cell-layers) corresponds to a CNN-layer (figure 4.1, top, middle). The properties of the different cell-layers (average diameter of the dendritic tree, temporal properties of the cell responds, etc.) can be approximated with appropriate CNN templates and parameters. The connections *between* these CNN layers (excitations, inhibitions, temporal delays, diffusion parameters, etc.) have also been defined in a way, so that they approximate the output of the corresponding retinal layers, as close as possible. The *temporal* properties of the retina channels have been entrapped with the adaptation of a weighted, circular memory buffer: the newly processed frame overwrites always the oldest, and the overall output of the given channel is the weighted, pixel-wise summation of the buffer content (figure 4.1, top of the image, right hand side).

The next step is the creation of the saliency maps belonging to the individual retina channels (figure 4.1, right hand side). These maps are being formed by differently sized receptive fields (RF). In other words, every channel has a different "optimal" receptive field size, or else, a different receptive field distribution (e.g. figure 6.11). In the beginning of the cerebral vision-processing (that is, in the "low" brain areas), the RFs are relatively small, and also circle-shaped. The higher we get in the brain hierarchy, the biggest the RFs are, concerning their size, and the more complex they become, with respect to their shape. Since in the beginning of my studies, the RF-sizes belonging to the different channels were practically unknown, in the initial state of the model, these have been adjustable values via the keyboard. (On figure 4.1, the parameters for which no literature data has been existent up to now are highlighted with red question marks.) The other important, yet unknown parameters which determine the final saliency map are the *weights* of the channel-based saliency maps (figure 4.1, bottom, middle).

I have approximated these parameters via human gaze direction measurements. For this purpose, I have applied an equipment called *"iView X Hi-Speed System"* suitable for gaze direction measurements. The "training-set", that is, the video clip that the subjects have seen for the process of *estimating* the parameters, was a ˜33 second flow, consisting of 267 frames, 8fps, where, each frame had a 512x298 pixel/frame resolution, 96 dpi. The stimulus did not contain any voice. It consisted of four shorter natural scenes, containing birds, mountains, lakes, horses, etc. The reason behind the usage of a *moving natural* input was justified by the fact that, according to literature-data, if the subject had no specific task to perform (e.g. into which continent the subject puts the scene, or, how many red and blue parrots the subject counts, etc.), then these conditions primarily trigger bottom-up visual attention.
During the measurements I have investigated the efficiency of 40 different receptive field sizes, for all the channels. This means, RF sizes spreading from 0.5° up to approximately 26°, expressed in terms of the viewing angle.
For the purpose of defining the *channel weights* I have applied different approaches addressing the following question: considering a given stimulus (frame), which channel(s) participate in triggering the saccade, and also, in what extent do these determine the new fixation position. (We call "saccades" those little eye-movements, "jumps", for which the center of the focus changes, that is, when one changes the fixation location. [1]) During the measurements I have applied 240 Hz sampling frequency and I have only taken into account the saccades bigger then 1°.

[1] In the literature we can find the word "saccade" in the sense of the shifting of the entire visual scene, but in the thesis I use this word in the sense given in the text.

For controlling the stimuli I have used the MatLab's Psychotoolbox[83], and for evaluating the measured data according to the above assumptions (differing according to which channels are being considered as saccade-triggering ones with respect to given stimuli) I have developed programs under MatLab as well.
For the purpose of *validating* the received parameters, I have performed similar human measurements on a "test video set" with an analogous topic (that is: moving natural scenes) using the same equipment. The other settings had been the same, but for the sake of accuracy, I have used a longer-duration stimulus including 9 scenes, 477 frames, ˜56 seconds.
During the validation process, I have measured the correspondence between the models' predictions and human gaze directions. For all the frames in the test video set, I have determined more points, as possible fixation locations (like: "on this frame, the coordinates of the most probable fixation location is the $x - y$ pair, the coordinates of the second most probable position is $x' - y'$ ", etc.). The results have shown a quite accurate correspondence: in ˜70% of the cases, the *measured* location was among the first four *predicted* locations. The accidental chance of this is less then 20%.

During the generation of a visual attentional model, the goal is to reproduce the accomplishment of living creatures, namely, the capability of finding the actually important visual information in the redundant and/or irrelevant torrent, in real time. This is possible by using heuristic methods and ideas as well, but the final goal is – primarily in the case of *neuromorphic* modeling – to understand and mimic the neural structure of the creature that we have used as model, as proper as possible. The importance of this lies on the fact that during the development of such a model, we can learn a lot about the functioning of living systems. Moreover, problems of an engineering design create a correlation loop with biological measurements as well. Furthermore, regarding efficiency, heuristic systems are hardly up to the operational level of the corresponding mechanisms in living creatures.

9.2 New scientific results

Thesis #1: *A new efficient method in the development of the bottom-up attentional model. The employment of the multi-channel mammalian retina model, which is based on the latest biological findings, instead of using the heuristic, low level visual feature filtering, and its' consequences.*

In living creatures, the information processing starts already in the retina. Even more, the information leaves the retina in a highly filtered and organized way, and projects towards the higher brain areas for further processing. The first precise enough neuro-biological descriptions of this information-classification – and thus also the retina-models built on them – have been appeared only in the last few years. Accordingly, this retinal process has been neglected in the earlier models, and instead of it, heuristic, different low level visual feature extraction algorithms have been applied.

The main novelties of the model I have implemented are the following: Firstly, the application of the methodology of the above mentioned multi-channel decomposition of the visual information, and thus the exploitation of the latest results of the retina research. Secondly, the estimation of the corresponding receptive field sizes in order to form the proper channel-based saliency maps by them.

1.1 I have improved the 'classical' visual attention model in a way that instead of using the generally applied 3-5 low level visual feature extraction (characterizing the 'classical' model), I am using the multi-channel mammalian retina decomposition method, which is based on the most recent neurobiological discoveries[21].

The first step in a neuromorphic visual attention model is the decomposition of the input image/video, according to the, so called, "low level visual features" (figure 4.1). Present-day models characteristically employ 3-5 of them, such as, edge-filtering, corner-filtering, color-filtering, etc.
Instead, in my model I have used the recently revealed and modeled mammalian

retina network model, which differentiates ten channels (figure A.2). In the case of five channels, the functions can be read of the output (edge-detection, motion-filtering, intensity and two color oppositions)[1], while the function of the remaining five is unknown, in the sense that, the aim of their process could not be formulated explicitly, at least up to present. Consequently, none of the heuristic models can incorporate them.

Corollary: In my model, similarly to living systems, the saliency maps that are based on those retina channels having non-explicitly described functions, also take part in the allocation of the fixation location. I have investigated their role on moving visual input.

The so called "saliency maps" are two dimensional, topographic maps of the physical world in the brain, such that, the activity of certain neurons are proportional with the 'vividness', 'high-contrast' of the corresponding locations in the physical world.

Since the retina channels having non-explicitly described functions form saliency maps as well, and thus they take part in the formation of the final saliency map, neglecting them significantly modifies the final results. In my model, I have taken into account the saliency maps for *all* the retina channels, and I have determined the weights, the 'importance' of the saliency maps belonging to these channels by the same method, that I have used for the explicitly formulated ones.

Seven channels' response (Transient, LED, Bistratified, Alpha, Beta, Delta, Polar) out of the ten, depend not only on the actual stimulus, but also on its' temporal behavior. In other words, the response of these channels – and accordingly the saliency maps based on them – more or less react on *changes*, on *motion*. The effect of these saliency maps, during the formation of bottom up visual attention, for the first time has been investigated during my measurements.

[1] According to the latest researches, certain cells in the retina respond to motion direction-dependently, that is, in certain living creatures, another channel could exist, which filters motion in a direction selective manner [61].

Thesis #2: *The estimation and optimization of the unknown parameters – namely, the receptive field sizes belonging to the different channels as well as the channel weights – based on human gaze-direction measurements. Additionally, the verification of the model, based also on human measurements.*

The model includes two essential, but unknown parameters: firstly, *what sized receptive fields* form the saliency maps on the different retina channels, and secondly, what is the *weighting* with which the channel-based saliency maps form the final saliency map. (These are marked with red question marks on figure 4.1) These parameters had been estimated via human gaze direction measurements, and I have checked the accuracy of the obtained model with similar measurements as well.

Directly, we can not measure the channel-based saliency maps (i.e. those belonging to a given retina channel) or their effects, but only the *fixation locations*, provided by the observers who have taken part in the experiments. We can only *infer*, deduce these immeasurable values by using different assumptions; that is, by using indirect methods. This is true for the *weighting* of the channels based maps as well. (It is quite difficult to design an experiment, a "stimulus", which affects only one of the channels – it is enough to mention, that if the stimulus is for example *dynamic*, then it immediately affects the seven spatio-temporal channels and the Intensity one as well.)
Since, according to literature data, the gaze directions controlled by bottom-up mechanism are essentially determined by these saliency maps, I have estimated their *efficiency* through their most intensive points, namely, via the correspondence between the 'keenest' locations of these channel-based saliency maps and the measured fixation locations. Consequently, I have estimated the missing parameters via *inferences* – which is another reason why the *validation* (comparison with human gaze direction measurements) has been so important.

II.1 I have determined optimal receptive field (RF) sizes for all the ten retina channels in our model, via human measurements. These correspond to those receptive fields sizes that generate the corresponding saliency maps. This process involves the investigation of 40 different RF sizes, between ˜0.5° and ˜26°, expressed in terms of the viewing angle.

For the same input, receptive fields with different sizes result in different saliency maps. I consider a receptive field size as *optimal*, if the saliency map created by it is the most *effective*, that is, for which the most intense points of the corresponding saliency map give the most accurate concurrence with the *measured* fixation locations.

Different saccades are provoked by different channels. The open questions are the following: ***1)*** *how many channels* take part in the provocation of a given saccade, and ***2)*** *which are these channels* concretely. Addressing these questions, I have investigated two different assumptions:

1. The channels which trigger a saccade (determine the new fixation location), are those being the most "effective" according to *arbitrary* receptive field sizes.
2. The channels which trigger a saccade are those that are effective *in average*, that is, all the saliency maps according to all the 40 receptive field sizes participate in the averaging.

I have investigated the results if the first 1, 3 and 5 most effective channels take part in the generation of the final saliency map, according to both assumptions. During the evaluation of the different cases, I have obtained curves similar to those that can be seen on figure 6.11. This diagram shows the curves that belong to the most accurate estimation.

For the different channels, the '*optimal*' receptive field sizes are those, by which the corresponding curves reach their maximum (tables 6.2 and 6.3). The

final, "tuned" model uses these RF sizes for creating the saliency maps.
In the 'real', *living* retina, the channels have an *interval* of RF sizes. In a biological viewpoint, the curves like figure 6.11 preferably show the *distribution* (density) of the different sized RFs, in the different retina channels. However, the explanation of the biological relevance of these curves was not the subject of my research. I emphasize that these investigations are based on a model level with aggregated functional tests and are not related to the neurobiological details.

II. 2. I have investigated different hypotheses addressing the question: what is the *proportion* ("weight") by which the different channels are responsible for provoking the saccades, that is, for determining the new fixation locations. Based on these, I have obtained different channel weightings.

I have analyzed assumptions, in which the channel weights had been kept constant, that is, the channel based-saliency maps had contributed in the formation of the final saliency map with always the same proportion. And also, I have investigated strategies, in which the channel weights had been constantly updated, according to the actual input.

- The ASSUMPTIONS FOR THE FIX CHANNEL-WEIGHTING STRATEGIES – which strongly build onto the previous point –, have been the following:
  The *channel-weights* are proportional to the relative *ratio* (percentage), by which they prove to be *saccade-triggering*:

  1. by *arbitrary* receptive field sizes
     (that is, how often do the highest saliency value(s) belong to the different channels - according to *any* RF)
     More concretely, a channel's weight is proportional to the frequency that the channel-based saliency map contained one of the highest values.
  2. by *average saliency value*
     (that is, how often do the different channels prove to be the most

salient one *in average* - using all the RF sizes)

More concretely, a channel's weight is proportional to the frequency that the channel-based saliency map was one of the highest in average.

The results are depicted on figure 6.6 and 6.12, whereas the accuracy of the different hypotheses can be seen on figure 7.1. On the diagrams, the first approach is denoted by "arf", whereas the second one by "avg".

- The hypothesis for determining the CHANNEL WEIGHTS IN A CONTINUALLY UPDATED MANNER is based on the assumption that the involvement of the different channels depend on the actual stimulus. In other words, the actual channel weights depend on the input, instead of being pre-defined.

  The two basic assumptions are the same than previously: those channel(s) are responsible for triggering a saccade on the actual stimulus, which:

  1. contains outstandingly high saliency values belonging to *any* RF size
  2. are the most salient *in average* on the given frame

  The weighting is proportional to these maximal/average values.

Contrary to the expectations – although the differences were small – the *fix* channel weighting strategies proved to be better than the *continually updated* ones, in the sense that they gave more accurate predictions, compared to human gaze direction measurements. The former strategies have performed better by ~5% than the latter ones (see figures 7.1 and 7.5).

Validation. I have verified the model's accuracy via human gaze direction measurements, and I have shown that the model predicts the human fixation locations with high conformity on complex natural scenes.

With the model adjusted according to the results of the described measurements, I have made predictions of the expected fixation locations, and then I have

compared them with measured human gaze directions. The *measured* locations were among the four most probable *predicted* locations in ˜70% of the cases, on the given frames (– the accurate value varies slightly according to the different hypothesis.) The accidental chance for this, under the same conditions, is a bit less than 20%. I have defined "*hit*", if the distance between the predicted and the measured location was less then 5° [1].
Figure 7.1 indicates the accuracy of the fix channel weighting strategies, figure 7.5 for the continually updated ones. On figure 7.1, the two approaches discussed in the text have been completed with a third one, in which the saliency map based on the Transient channel (which in-filters everything that moves and eliminates all the steady part) forms the final saliency map, on its' own. According to literature-data on dynamic input, this channel is outstandingly strong – which is intuitively not surprising, if we take into account how naturally we snap our head at cats, birds, etc., if they abruptly make a motion on the periphery of our sight. These results have been confirmed by my measurements as well (left-most columns in the bar-trios).

9.3 Applications of the results

Areas where attentional models can be applied are extremely wide, the subtasks and methods employed within them can be used in very many fields. Accordingly, during the last years, I have had the opportunity to test different parts of my model in real practical applications as well – namely in the "*Bionic Eyeglass Project*".

This project meant to help the everyday life of blind or visually impaired people with mobile equipment, via image-flow analysis and different recognition methods. The main lines, like the subtasks, have been developed together with the expert of the "Hungarian National Association of Blind and Visually Impaired

[1] Counting with 10°, the hit ratio ameliorates significantly – although of course the accidental chance as well. The 5° 'threshold' seemed to be a reasonable choice, both from biological and from evaluational viewpoints.

People". Within this project, I have successfully adapted different parts of the discussed model, or rather, of an expanded version of it. This version includes a *preprocessing part designed to stabilize the unstable input* that comes from a camera held by a blind walking person. These video-flows are usually extremely noisy and unstable, often accompanied by fast and unexpected camera motions. Additionally, often the picture's main objects shift significantly from one frame to another, e.g. during turning around. The goal of the image stabilization step is to keep the steady objects (e.g. buildings) in the same pixel positions, while the moving objects (for example the pedestrians) can change position.
The main idea in this step is to combine an optic flow algorithm with an affine transformation model, which can handle translation, scaling, rotation and shear. By using the optic flow algorithm we obtain estimation for the velocity of the pixels by measuring their time and spatial gradients (vertical and horizontal) piece by piece. Then, with the transformation model, the translation (in vertical and horizontal directions), scaling, rotation and shear *of the frame* can be estimated. By the usage of the mammalian retina channel decomposition, the classical difficulty that image processing algorithms nowadays face (namely that the intensity or color values of the same object largely depend on the actual lighting conditions) can be avoided – at least partly. This observation has a fundamental importance in practical applications, and it is exploited in the methods aiming to solve the following problems raised within the Bionic Eyeglass Project:

- Locating LED indicators (in real-life indoor and outdoor scenes)

- Finding traffic signs in real-life street scenes
  The main purpose of these two tasks is to realize a fast method that locates the areas which contain the traffic signs / LED indicators with high probability, on complex real-life outdoor scenes. Subsequently, a classifier algorithm has to analyze only the located ROIs ("Region of Interest") instead of the whole input, which can fasten up the whole process significantly. The main difficulties derive from the instability of the by-default bad-resolution input, the unconstrained lighting conditions, and from the *variety* of the possible inputs.
  The accuracy of the introduced methods is around 80%. The test database

has been made out of complex real-life scenes, for all the different tasks.

- Finding light sources (lamps) – which task (although it seems to be a trivial 'problem' for a person with normal vision), could prevent annoyance for visually impaired people, for example, by preventing the lamps to remain switched-on for weeks after a guest.
  Here, the most important criterion is that the solution has to be independent from the input's actual brightness, that is, the accuracy should be the same in the case of a sun-drenched room and a dark cell.
  The method I have introduced relies only on a single retina channel, the "Polar" channel, and achieves a very high accuracy: the ratio of the correct answers is around 99%.

The precise algorithms have been explained in chapter 8 and have appeared in separate publications.

Generally speaking, the possible application-fields of a well functioning visual attentional system is extremely wide, starting from different monitoring systems via robot vision up to different 'bionic' applications. Nevertheless, a well functioning *bottom-up* system (which I have attempted to produce during my Ph.D. studies) is not a *complete* attentional system. It would be complete, if it had included the so called "top-down" method as well. However, our knowledge of this cortex-originated function is quite restricted for the time being, but at any rate, slimmer than necessary for a reliable and complete model.
At the same time, regarding the above task, *some* knowledge we already possess comes from well known data from the literature, for example, that this method is "fed-back" at the point of summing up the channel-based saliency maps, right before the creation of the final saliency map (figure 4.1, bottom, middle). On this schema, different practical applications can be constructed, for example via the task-dependent modification of these weights (e.g. finding traffic signs, from the above discussed applications).

Acknowledgements

First of all, I would like to thank Professor Tamás Roska, my supervisor, for the vast encouragement, guidance and help, for the stirring discussions and consultations, and, because he could always find time for me in his extremely dense time-schedule.
Similarly, I would like to say thanks for my consultant, Dr. Zoltán Vidnyánszky for his concern about me, for the pieces of advice, for the discussions, the careful reviewing of my publications, and also, that via him I could have a blink into the world and thinking of the biologists.

I would also like to thank my fellows at the Doctoral School for the discussions and assistance concerning a wide range of topics. To István Kóbor, László Havasi, Tamás Bárdi, Dániel Hillier and Gábor Vásárhelyi, with whom I have been sharing a common workplace, and moreover to György Cserey, Gergő Soós, Barna Hegyi, Tamás Harczos, Csaba Benedek, Éva Bankó and Viktor Gál, without whom these years would have been much less exciting.

Also, special thanks for Dávid Bálya, for the help in the retina model I have used during my doctoral studies, and for the guidance he gave, primarily during my first year.
Similarly, I would like to express my thanks to the "vision", or rather "Bionic Eyeglass" group for the co-operation: to Dávid Bálya, Róbert Wagner and to Kristóf Karacs.

I am grateful for the help and good advice I have got from Professor György Karmos before my stay in Leuven: it would had been more difficult without them!
I am particularly indebted to Professor Marc Van Hulle as well, who has made possible an interesting and pleasant semester in his laboratory in Leuven. And also, from his group to Karl Pauwels, for his persistent help, and because he always had time for me, even for hours.

I keep beautiful memories of my university years which I had the luck to spend in Szeged. From there, I would like to thank András Kocsor, my former supervisor for guiding my first efforts, my first paper.

And last but not least, I am very-very grateful to my family, to David for being a great and unique brother, and to my parents for their endless patience, love and support, which I have had the good luck to enjoy during all my studies. And also, to Elias, for his constant understanding and help in these “dense” periods.

Appendix A

The used multi-channel retina simulator

The retina channel simulator I have used relies on the retina-model developed by David Balya and which is detailed in [9]. I have prepared both the used CNN (Cellular Neural/Nonlinear Network, [84]) simulator and the multi-channel retina model (similarly to the entire attentional model) in Borland C++, while the proper parameters, which define the exact spatio-temporal behavior of the different retina channels have been given by David Balya. This section is devoted to summarize the used retina model, but more details can be found in [9].

The usage of Cellular Neural/Nonlinear Network (CNN)-based algorithms in handling different visual problems is common: from robot navigation [85] to motion analysis [86] the range is wide. The retina model I have used to perform the seven spatio-temporal channels (Transient, Local Edge Detector (LED), Bistratified, Alpha, Beta, Delta and Polar, figure A.2.) is also CNN-based and has been developed by keeping the main structure of the retina in a manageable simple form. The circuit structure of the mammalian retina and its' multilayer spatial temporal model is the same [87, 9], (figure A.1).

The sketch of a (general) spatio-temporal channel is depicted on figure A.1,a. Each horizontal line on the right-hand side is a CNN layer which corresponds to a retina layer (depicted in the left-hand side of the picture). The outer retina, which is the same for all the channels, consists of the cone and the horizontal layer. The horizontal layer feeds back to the cone layer through an inhibitory

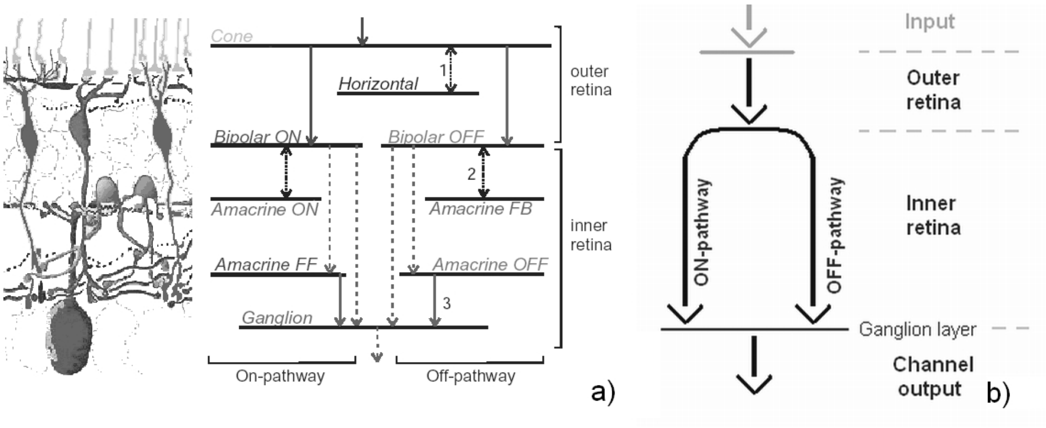

Figure A.1: The scheme of a general retina channel (b) roughly and (a) with CNN layers. In our model we have seven of these, one for each ganglion output. The interacting diffusion layers are numbered. The dashed lines show the inhibitory connections whereas the continual ones nominate the excitatory ones. Figure (a) is cited from [9].

connection; thus, the output of the cone layer includes the effect of the horizontal cells as well.

The bipolar cells connect the inner and the outer retina. From an engineering viewpoint the inner retina can be divided into an On- and an Off-pathway. (figure A.1,b) “On” cells respond during illumination, “Off” cells respond when the light disappears, whereas “On-Off” cells react on both cases.

Each channel consists of three layer pairs, which are serially connected. The first one is the cone-horizontal, which composes the outer retina. The second one is the amacrine-bipolar, where the connection is also inhibitory similarly to the previous one. The third connection is excitatory between the amacrine and the ganglion layers. The output of the retina-channel is the output of the ganglion layer. Ganglion cells typically have two qualitatively different inputs: an excitatory and an inhibitory one. Excitation comes from the amacrine layer, whereas inhibition derives from the bipolar cells.

For the seven spatio-temporal channels (which differ only in the parameters that determine their spatio-temporal characteristics), firstly I have performed the *temporal processing*. For this purpose I have used a buffer for the images,

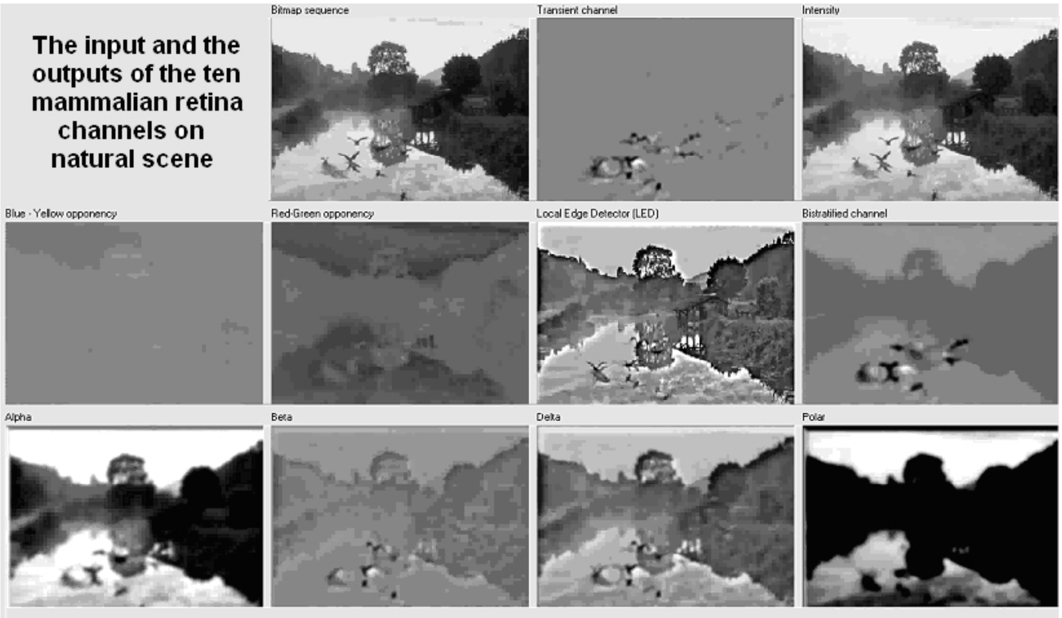

Figure A.2: An example for the functioning of the retina. The input image (first picture) is processed by ten different pathways resulting in ten ganglion-cell types that form the ten retina channels. The second picture in the first row (next to the input image) is the output of the 'Transient' channel that filters out the mobile parts of the visual scene and removes all the steady sections: at this moment the birds flight triggers the biggest response. Normally this is one of the strongest channels. The last image in the first row depicts the output of the 'Intensity' channel. In the second row we can see the blue-yellow and the red-green contrast channels (these are the color channels), the LED (local edge detector) and the 'bistratified' channels. The functions of the channels depicted in the third row (Alpha, Beta, Delta and Polar) are unknown for the present, as well as the bistratified channel's task. The picture is from [8].

which preserved the recently processed sceneries – in the biological equivalent this corresponds to the information that is still under processing in deeper layers of the retina. Practically, this is a fixed-sized buffer, where the certain positions indicate the time elapsed since the input reached the sensor. Each of these positions has different weights. (It is important to note that working with image frames is a corollary of working with simulators that run on PCs; this is because of the fact that the retina has no frame-rate or any similar category: it works on a totally analog way, in the sense that the input image flow is continuous in time and value, and there is no time discretization, the only discretization is in space.)

Once a new frame is being read, it gets diffused with the former images: that is, the signals that reach the retina beforehand, subsequently reside on different levels of the vertical pathway. The different layers of the retina have different diffusion characteristics: accordingly, the individual positions of the circular buffer have different weights that characterize the diffusion being made on the image restored there. Once this process has been completed, the result overwrites the oldest image. This is the outcome of the specific retina channel.
Spatial processing is the effect of the diffusions that occur inside the certain layers. From an engineering viewpoint this is the outcome of the subtraction being made between two different diffusions engaged on the last (temporally already processed) frame. Figure A.2 shows a snapshot of the ten retina channels for a natural scene.

The remaining three channels (red-green color-opposition, blue-yellow color-opposition, and Intensity) do not require complex simulations, since they use only *actual data* for producing the output. Although some basics are known [20], the precise method explaining how the colors are processed is mostly undiscovered. As a wildly accepted approach, if 'R' is the actual red value in a given pixel-position (from the 'RGB' triplet), 'G' is the green and 'B' is the blue, then [20]

- Intensity has been calculated as $0.812 * G + 0.177 * R + 0.1 * B$
- Red-green opposition as $R - G$
- Blue-yellow opposition as $B - \frac{R+G}{2}$

Appendix B

The Set-Up of the Measurements

B.1 The Stimuli

I prepared two different video sets. One of them was used for the basic measurements, the final purpose of which was to estimate the unknown parameters, namely the (most effective) RECEPTIVE FIELD SIZES, and the CHANNELS WEIGHTS (section 4.2). The second video was used for the validation and assessment of the model.
Both video-sets contained moving natural scenes, each without any humans or artificial environment: birds, horses, rivers flowers, sees, mountains, etc. The stimulus was 8 frame/second video, 512x298 pixel/frame, 96 dpi each. No audio was added. The first ("training") video-set included 4 clippets, 267 frames, ˜33 seconds. The validation (or "test" video-) set contained 9 clippets with a sum of 447 frames, ˜56 seconds. Participants were asked to watch both videos 4 times in the following order: 2 for the training video, then 2 for the test video, then 2 for the training video again, and finally 2 for the test one.

The reason why I chose *natural* scenes is because recent results indicate that under natural viewing conditions (if the subjects have no specific task to perform, for example counting the birds, making suggestions where the scene could be, etc.) attention is indeed guided by bottom-up mechanisms [62]. Secondly, the reason why I chose *moving* stimulus, is because as it is detailed in the section dealing with the retina channels (sections 2.2.1 and 3.2), seven channels out of the ten have some kind of “memory”, thus they give fundamentally different response on steady input then on moving stimuli.

B.2 Participants

21 naive human observers participated in the first “training” video-set measurements (and 2 non-naive) and 14 naive (plus 1 non-naive) in the second series. Non-naive participant’s data were not included in the evaluation. Each subject had normal or corrected-to-normal vision.

B.3 Experimental Design

The equipment I used for recording the fixation locations was an “iView X Hi-Speed System” which I used on 240 Hz sampling frequency. The distance between the subject’s eyes and the monitor was 50 cm; the inner part of the monitor was 40 cm x 30 cm. I recorded saccade-end locations and in the first series I processed the data belonging to saccades bigger then 1 degree, in order to find out the most BU-modified fixations. During 66 (naive) trials in the first case I recorded 3995 fixations, from which, 2560 saccades were bigger then 1 degree. The second run, for validation, included 54 trials with 6430 saccade end-location recordings.

Appendix C

Used Definitions, Appellations, Abbreviations

- **Attention** is the cognitive process of selectively concentrating on one aspect or feature of the environment while ignoring the other ones. It can belong to any sensory system: audio, visual, tactile or smelling.

- The **Bottom-Up** (OR BY ABBREVIATION: "**BU**") METHOD is an important mode of operation of attention, which is largely unconscious ("reflex-like") and driven by the specific attributes of the stimuli present in the visual environment [14]. The "bottom-up" and the "top-down" attentional mechanisms form the entire attentional process in conjunction with each other.

- **Channel-based saliency map (Feature-dependent saliency map)** is a scalar, two-dimensional map whose activity topographically represents the visual saliency of that particular visual feature, which the given channel codes. That is, a red object in a field of green objects is only salient in the saliency map belonging to a channel that codes red-green opposition, and will not cause activation in other maps.

- A **CNN (Cellular Neural/Nonlinear Network):** is basically a coupled, dynamic, analog, non-linear processor-array, which lies on an $M \times N$ tetragonal grid, consisting of locally interconnected cells. There is a cell (a dynamic system, which can be a processor) in each node of the grid, which is connected to only the surrounding cells in an r radius distance. It can include more of these layers (see figure). The *state equation* which describes the state if the cell indexed with (i, j), is the next [84]:

$$\dot{x}_{i,j} = -x_{i,j} + \sum_{C(k,l)\in S_r(i,j)} A\left(i,j;k,l\right) y_{k,l} + \sum_{C(k,l)\in S_r(i,j)} B\left(i,j;k,l\right) u_{k,l} + z_{i,j}$$

where $x_{i,j} \in R$, $y_{k,l} \in R$, $u_{k,l} \in R$ and $z_{i,j} \in R$ are called the **state**, **output**, **input**, and **threshold** of cell $C\left(i,j\right)$, respectively. $A\left(i,j;k,l\right)$ and $B\left(i,j;k,l\right)$ are called the **feedback** and the **input synaptic** operators (*"templates"*). The output of the cell $C_{i,j}$ is:

$$y_{i,j} = f\left(x_{i,j}\right) = \frac{1}{2}\left|x_{i,j}+1\right| - \frac{1}{2}\left|x_{i,j}-1\right|$$

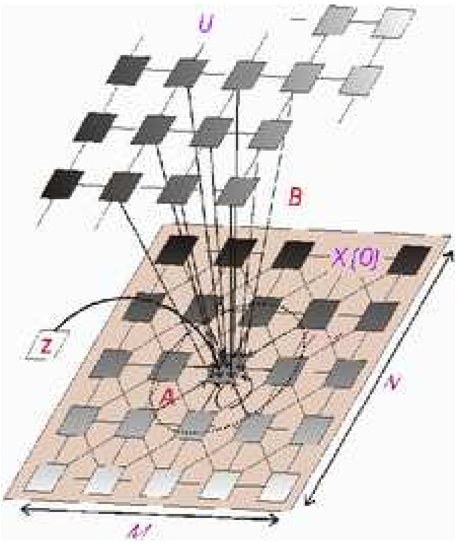

Figure C.1: The basic structure of a Cellular Neural/Nonlinear Network. The picture is from [10].

- **“Efficiency” of a [retina channel – receptive field] pair**: A receptive field (RF) is “*effective*” on a given channel, if the saliency values determined by it reaches their maximal values on those locations where human observers attend to with a good chance.

- **Fixation location**: That small part of the visual environment which is mapped into the fovea, to the center of sharp vision.

- **Low-level visual features**: Basic characteristics of a visual stimulus, for example the edges, junctions, intensity, color-properties, moving parts, orientation, etc. of its’ elements.

- **Master** (OR **final**) **saliency map** is a scalar, two-dimensional map whose activity topographically represents visual saliency, irrespective of the feature dimension that makes the location salient. That is, an active location in the saliency map encodes the fact that this location is salient, no matter whether it corresponds to a red object in a field of green objects, or to a stimulus moving towards the right while others move towards the left [6].

- **Naive subject**: Human experimental subject without any information or knowledge about the set-up or goal of the given experiment which could effect her behavior.

- **Optimal RF**: The most effective receptive field size on a given channel (see above “efficiency”).

- The **retina** is the light sensitive inner layer of the *eye* consisting of neurons, which receives images formed by the lens and transmits them to the brain through the optic nerve. The optic nerve is formed by the axons of the so called ganglion cells.

- **Retina channels:** The retina not only captures the visual stimulus and conveys it to the brain, but it already starts to *process* it: the visual input dissolves according to around dozen different "low-level visual features", like motion, color-oppositions, edge detections, etc. These are the so called RETINA CHANNELS, which thus code different aspects of the visual environment. (see figure 3.3, section 3.2)

- **RF: Receptive field** of a cell: That area of retina over which light stimuli changes the activity of a particular cell [35]. More general, every receptor organ and cell has a receptive field, a specific part of the world to which it responds [4].

- **Saccades** are those little eye-movements, "jumps", for which the center of the focus changes, that is, when one changes the fixation location. (The word probably originates from the 1880s, when French ophthalmologist Émile Javal used a mirror on one side of a page to observe eye movement in silent reading, and found that it involves a succession of discontinuous individual movements.)

- The **Saliency map** (OR **Saliency matrix**) is a scalar, two-dimensional map whose activity topographically represents visual saliency. The channel-based (or feature dependent) saliency maps form the final (or) master saliency map (see above).

- **Top-down** (OR BY ABBREVIATION: "**TD**") METHOD is the volitional mode of operation of attention, which is largely determined by the current goals and state of the organism. The "bottom-up" and the "top-down" attentional mechanisms form the entire attentional process in conjunction with each other.

- **Trial**: a scientifically controlled study of some kind of behavior in well-defined conditions. In this case the observed behavior has been the visual attention (gaze direction) under bottom-up conditions, using human subjects.

- **Visual Attention**: Selective visual attention is the mechanism by which we can rapidly direct our gaze towards objects of interest in our visual environment [14].

References

[1] http://www.amdcanada.com/. (document), 2.1

[2] http://www.owlnet.rice.edu/ psyc351/Images/RetinaLayers.jpg. (document), 2.2

[3] http://www.brainconnection.com/topics/?main=anat/receptive. (document), 2.3

[4] B. Kolb and I. Q. Whishaw, *Fundamentals of Human Neuropsychology.* W.H.Freeman & Co Ltd, 2003. (document), 2.2.3, 2.5, C

[5] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985. (document), 3, 3.1, 5

[6] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Neuroscience*, vol. 2, pp. 1–11, 2001. (document), 1.1, 1.2, 2, 3, 3.1, 3.1, 6, 6.4, C

[7] A. K. Lázár, R. Wagner, D. Bálya, and T. Roska, "Functional representations of retina channels via the refinec retina simulator," in *Cellular Neural Networks and their Applications. Proceedings of the 8th IEEE international workshop Budapest*, pp. 333–338, 2004. (document), 3.2, 3.2, 3.3

[8] A. Lazar, Z. Vidnyanszky, and T. Roska, "Modeling stimulus-driven attentional selection in dynamic natural scenes," *International Journal on Circuit Theory and Applications.* (document), 5.2, A.2

[9] D. Balya, B. Roska, T. Roska, and F. Werblin, “A cnn framework for modeling parallel processing in a mammalian retina,” *International Journal on Circuit Theory and Applications*, vol. 30, pp. 363–393, 2002. (document), 8.3.2, A, A.1

[10] http://lab.analogic.sztaki.hu/. (document), C.1

[11] J. K. Tsotsos, L. Itti, and G. Rees, “A brief and selective history of attention,” in *Neurobiology of Attention*, pp. xxiii–xxxii, 2005. 1.1

[12] R. Desimone and J. Duncan, “Neural mechanisms of selective visual attention,” *Annu. Rev. Neurosci.*, vol. 18, pp. 193–222, 1995. 1.1

[13] F. Crick and C. Koch, “Constraints on cortical and thalamic projections: the no-strong-loops hypothesis,” *Nature*, vol. 391, pp. 245–250, 1998. 1.1

[14] L. Itti, “Modeling primate visual attention,” in *Computational Neuroscience: A Comprehensive Approach (J. Feng Ed.)*, pp. 635–655, 2003. 1.1, 2, 6, C, C

[15] J. R. Bergen and B. Julesz, “Parallel versus serial processing in rapid pattern discrimination,” *Nature*, vol. 303, pp. 696–698, 1983. 1.1

[16] K. Nakayama and M. Mackeben, “Sustained and transient components of focal visual attention,” *Vision Research*, vol. 29, pp. 1631–1647, 1989. 1.1

[17] J. Braun and D. Sagi, “Vision outside the focus of attention,” *Percept. Psychophys*, vol. 48, pp. 45–58, 1990. 1.1

[18] O. Hikosaka, S. Miyauchi, and S. Shimojo, “Orienting a spatial attention – its reflexive, compensatory, and voluntary mechanisms,” *Brain research*, vol. 5, pp. 1–9, 1996. 1.1

[19] J. Braun and B. Julesz, “Withdrawing attention at little or no cost: detection and discrimination tasks,” *Percept. Psychophys*, vol. 60, pp. 1–23, 1998. 1.1

[20] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*. Appleton&Lange, 3 ed. 1.2, 2.2.1, 2.2.3, 3.2, A

[21] B. Roska and F. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature*, vol. 410, pp. 583–587, 2001. 1.2, 2.2.1, 3.2, 3.2, 8.3.2, 9.2

[22] B. Gulyás, G. Kovács, and Z. Vidnyánszky, "Consciousness and cognitive neurosciences," in *Cognitive Neuroscience*, pp. 619–649, 2003. 2

[23] R. Parasuraman, "The attentive brain: issues and prospects," in *The Attentive Brain*, pp. 3–15, MIT Press, 1998. 2

[24] Z. Vidnyanszky, "Visual attention," in *Cognitive Neuroscience*, pp. 219–234, 2003. 2, 8.5

[25] J. B. Hopfinger, M. H. Buonocore, and G. R. Mangun, "The neural mechanisms of top-down attentional control," *Nature Neuroscience*, vol. 3, pp. 284–291, 2000. 2.1

[26] M. Corbetta, J. M. Kincade, J. M. Ollinger, M. P. McAvoy, and G. L. Shulman, "Voluntary orienting is dissociated from target detection in human posterior parietal cortex," *Nature Neuroscience*, vol. 3, pp. 292–297, 2000. 2.1

[27] B. C. Motter, "Neural correlates of attentive selection for color or luminance in extrastriate area v4.," *J. Neurosci.*, vol. 14, pp. 2178–2189, 1994. 2.1

[28] S. Treue and J. C. M. Trujillo, "Feature-based attention influences motion processing gain in macaque visual cortex," *Nature*, vol. 399, pp. 575–579, 1999. 2.1, 3.1

[29] F. Barcelo, S. Suwazono, and R. T. Knight, "Prefrontal modulation of visual processing in humans," *Nature Neurosci*, vol. 3, pp. 399–403, 2000. 2.1

[30] R. H. Masland, "The fundamental plan of the retina," *Nature neuroscience*, vol. 4(9), pp. 877–886, 2001. 2.2.1, 2.2.1, 3.2

[31] R. Sekuler and R. Blake, *Perception.* McGraw-Hill,Inc., 3 ed., 1994. 2.2.2

[32] K. Suder and F. Worgotter, “The control of low-level information flow in the visual system,” *Rev. Neurosci.*, vol. 11, pp. 127–146, 2000. 2.2.2

[33] J. E. Hummel and I. Biederman, “Dynamic binding in a neural network for shape recognition,” *Psychol. Rev.*, vol. 99, pp. 480–517, 1992. 2.2.2

[34] J. H. Reynolds and R. Desimone, “The role of neural mechanisms of attention in solving the binding problem,” *Neuron*, vol. 24, pp. 19–29, 1999. 2.2.2

[35] T. Vilis, “The physiology of the senses.” http://www.physpharm.fmd.uwo.ca/undergrad/sensesweb. 2.2.3, C

[36] A. Treisman and G. Galade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, pp. 97–136, 1980. 3

[37] W. Osberger and A. Rohaly, “Automatic detection of regions of interest in complex video sequences,” *Human Vision and Electronic Imaging VI, Proceedings of SPIE*, vol. 4299, 2001. 3, 6.4

[38] L. Itti, “Models of bottom-up attention and saliency,” pp. 576–582. 3, 3.1

[39] J. Khan and O. Komogortsev, “A hybrid scheme for perceptual object window design with joint scene analysis and eye-gaze tracking for media encoding based on perceptual attention,” *SPIE Journal of Electronic Imaging*, vol. 15(2), 2006. 3

[40] G. Deco, E. T. Rolls, and J. Zihl, “A neurodynamical model of visual attention,” in *Neurobiology of Attention*, pp. 593–599, 2005. 3

[41] S. Shipp, “The brain circuitry of attention,” *Trends in Cognitive Sciences*, vol. 8(5), 2004. 3

[42] R. Carmi and L. Itti, “Visual causes versus correlates of attentional selection in dynamic scenes,” 2006. DOI:10.1016/j.visres.2006.08.019. 3, 6.4, 7.1

[43] R. Carmi and L. Itti, “The role of memory in guiding attention during natural vision,” *Journal of Vision*, vol. 6(9), pp. 898–914, 2006. 3, 5, 6.4

[44] J. M. Wolfe, "Guidance by visual search by preattentive information," in *Neurobiology of Attention*, pp. 101–104, 2005. 3.1

[45] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electronic Imaging*, vol. 10(1), pp. 161–169, 2001. 3.1

[46] H. C. Nothdurft, "Texture discrimination by cells in the cat lateral geniculate nucleus," *Exp. Brain Res.*, vol. 82, pp. 48–66, 1990. 3.1

[47] J. Allman, F. Miezin, and E. McGuinness, "Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons," *Annu. Rev. Neurosci.*, vol. 8, pp. 407–430, 1985. 3.1

[48] J. M. Wolfe, "Visual search in continuous, naturalistic stimuli," *Vision Research*, vol. 34, pp. 1187–1195, 1994. 3.1

[49] M. M. Chun, "Contextual guidance of visual attention," in *Neurobiology of Attention*, pp. 246–250, 2005. 3.1

[50] J. Braun, "Vision and attention: the role of training," *Nature*, vol. 393, pp. 424–425, 1998. 3.1

[51] M. Ahissar and S. Hochstein, "The spread of attention and learning in feature search: effects of target distribution and task difficulty," *Vision Research*, vol. 40, pp. 1349–1364, 2000. 3.1

[52] M. Sigman and C. D. Gilbert, "Learning to find a shape," *Nature Neuroscience*, vol. 3, pp. 264–269, 2000. 3.1

[53] V. Navalpakkam, M. A. Arbib, and L. Itti, "Attention and scene understanding," in *Neurobiology of Attention*, pp. 197–203, Elsevier, 2005. 3.1, 6

[54] G. Deco and J. Zihl, "A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system," *Journal of Computational Neuroscience*, vol. 10(3), pp. 231–253, 2001. 3.1

[55] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, pp. 1019–1025, 1999. 3.1

[56] M. Riesenhuber and T. Poggio, "Models of object recognition," *Nature Neuroscience*, vol. S3, pp. 1199–1204, 2000. 3.1

[57] K. M. O'Craven, P. E. Downing, and N. Kanwisher, "fmri evidence for objects as the units of attentional selection," *Nature*, vol. 401, pp. 584–587, 1999. 3.1

[58] P. R. Roelfsema, V. A. Lamme, and H. Spekreijse, "Object based attention in the primary visual cortex of the macaque monkey," *Nature*, vol. 395, pp. 376–381, 1998. 3.1

[59] R. A. Abrams and M. B. Law, "Object-based visual attention with endogenous orienting," *Percept. Psychophys.*, vol. 62, pp. 818–833, 2000. 3.1

[60] R. M. Klein, "Inhibition of return," *Trends in Cognitive Science*, vol. 4, pp. 138–147, 2000. 3.1

[61] S. I. Fried, T. A. Muench, and F. S. Werblin, "Mechanisms and circuitry underlying direction selectivity, in the retina.," *Nature*, vol. 420, pp. 411–414, 2002. 1, 1

[62] D. J. Parkhurst and E. Niebur, "Stimulus-driven guidance of visual attention in natural scenes," pp. 240–245, 2005. 5, 6, 6.4, B.1

[63] C. M. Privitera and L. W. Stark, "Scanpath theory, attention, and image processing algorithms for predicting human eye fixations," in *Neurobiology of Attention*, pp. 296–299, Elsevier, 2005. 6

[64] C. Zetzsche, "Natural scene statistics and salient visual features," in *Neurobiology of Attention*, pp. 226–232, Elsevier, 2005. 6, 6.4

[65] D. J. Parkhurst and E. Niebur, "Texture contrast attracts overt visual attention in natural scenes," *European Journal of Neuroscience*, vol. 19, pp. 783–789, 2004. 6, 6.4

[66] S. A. Adler, "Visual search and popout an infancy," in *Neurobiology of Attention*, pp. 207–212, 2005. 6

[67] C. Zetzsche, "Natural scene statistics and salient visual features," in *Neurobiology of Attention*, pp. 226–232, 2005. 6

[68] J. Theeuwes, "Irrelevant singletons capture attention," in *Neurobiology of Attention*, pp. 418–427, 2005. 6

[69] H.-C. Nothdurft, "Salience of feature contrast," in *Neurobiology of Attention*, pp. 233–239, 2005. 6

[70] M. A. Lewis, "The role of visual attention in the control of locomotion," in *Neurobiology of Attention*, pp. 638–641, 2005. 8

[71] L. Paletta, E. Rome, and H. Buxton, "Attention architectures for machine vision and mobile robots," in *Neurobiology of Attention*, pp. 642–648, 2005. 8

[72] H. Yee and S. Pattanaik, "Attention for computer graphics rendering," in *Neurobiology of Attention*, pp. 649–651, 2005. 8

[73] G. Medioni and P. Mordohai, "Saliency in computer vision," in *Neurobiology of Attention*, pp. 583–585, 2005. 8

[74] T. Roska, K. Karacs, R. Wagner, A. Lazar, D. Balya, and M. Szuhaj, "Bionic eyeglass: an audio guide for visually impaired," *Proceedings of the 1st Biomedical Circuit and System Conference, London*, pp. 190–193, 2006. 8

[75] T. Roska, D. Balya, A. Lazar, K. Karacs, R. Wagner, and M. Szuhaj, "System aspects of a bionic eyeglass," *IEEE international symposium on circuits and systems. Island of Kos*, pp. 161–164, 2006. 8

[76] A. Lazar, K. Pauwels, M. Van Hulle, and T. Roska, "Scene analysis of unstable video flows – using multiple retina channels and attentional methods," in *Integrated Circuits: Research, Technology and Applications*, NovaScience, 2008. 8

[77] B. Zitova and J. Flusser, "Image registration: a survey," *Image and Vision Comp.*, vol. 21, pp. 977–1000, 2003. 8.1

[78] M. Otte and H. H. Nagel, "Optical flow estimation: Advances and comparisons," *Computer Vision, ECCV-94*, vol. 800, pp. 51–60, 1994. 8.1

[79] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–204, 1981. 8.1.1

[80] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12(1), pp. 43–77, 1994. 8.1.2

[81] Mihaly Szuhaj, personal communication. 8.3.2

[82] B. Gulyas, "Functional neuroimaging in cognitive neurosciences," in *Cognitive Neuroscience*, pp. 103–125, 2003. 8.5

[83] MatLab Version 7.1.0.246 (R14) Service Pack3. 9.1

[84] L. O. Chua and T. Roska, *Cellular Neural Networks and Visual Computing.* Cambridge University Press, 2002. A, C

[85] X. Vilasís-Cardona, S. Luegno, J. Solsona, A. Maraschini, G. Apicella, and M. Balsi, "Guiding a mobile robot with cellular neural networks," *International Journal on Circuit Theory and Applications*, vol. 30, pp. 611–624, 2002. A

[86] B. Shi, "An eight layer cellular neural network for spatio-temporal image filtering," *International Journal on Circuit Theory and Applications*, vol. 34, pp. 141–164, 2006. A

[87] F. S. Werblin, T. Roska, and L. O. Chua, "The analogic cellular neural network as a bionic eye," *International Journal on Circuit Theory and Applications*, vol. 23, pp. 541–569, 1995. A