POLITECNICO
DI TORINO

Master's Degree in Biomedical Engineering

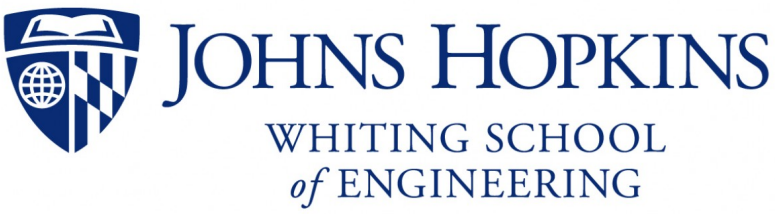
Master's Degree Thesis

Computational Stereo-Vision Model of Proto-Object based Saliency in Three-Dimensional Space

Advisor
Danilo Demarchi
Co-Advisor
Ralph Etienne-Cummings

Candidate
Elena Mancinelli

December 2018



The study presented in this dissertation was entirely carried out at the Computational Sensory-Motor Systems Laboratory of the Electrical and Computer Engineering Department at Johns Hopkins University Whiting School of Engineering, Baltimore, Maryland, USA, under the supervision of professor Ralph Etienne-Cummings, chairman of the department. This gave me direct access to a large amount of invaluable material, fundamental for the completion of the study. All the results were discussed with Ernst Niebur from Solomon Snyder Department of Neuroscience, Johns Hopkins University, who, thanks to his decades of experience in the field, provided precious advice.

We think too much and feel too little.
More than machinery, we need humanity;
more than cleverness, we need kindness and gentleness.
Without these qualities, life will be violent and all will
be lost.

Charlie Chaplin

Acknowledgments

First of all I want to say that these pages are much more than a research summary. These pages are an overseas journey through a state, the United States of America, that until only few months ago seemed so hard to reach.

Having the possibility to confront myself with the American scientific research was something that I have been dreaming about since my first university years. Now I can say that, thank to the Politecnico di Torino and above all to the kindness and willingness to share of professor Danilo Demarchi, my dreams have become true.

Thank you professor because without your constant support and your overwhelming passion in doing your job, I could not have lived one of the most intellectually stimulating period of my life.

Finding yourself in such prestigious university reality as the Johns Hopkins University in Baltimore was something that terrified me, but it is thanks to people like Professor Ralph-etienne Cummings that fears can turn into small everyday challenges to be faced and overcome with great satisfaction. Thank you Professor Ralph for stimulating my desire for knowledge in a constant manner and without ever preferring the role of the head to that of guidance.

A big thank you to professor Ernst Niebur for all the precious advice and the inestimable knowledge made available to me.

Thank you to all the staff of the ECE department, Nicole, Melissa, Cora, Makea, Eileen, Debbie and Dana. Your welcome, your hugs, your kindness, your smile and your coffees have been a sweet morning appointment during my entire period at JHU.

I want to thank all my friends and colleagues form CSMS Lab, John, Adam K., Adam C., Bayo, Takeshi, Duncan, Patrick, Tao and Jamal for being a constant source of inspiration, for having welcomed me with esteem and naturalness. Thank you because you gave me the certainty of having a team at 6500 kilometers from home.

A big thank you to all the Fosters' Lab, starting from Amy and Mark that giving me a desk in their lab, gave me the invaluable gift of having the opportunity to get closer to the fascinating field of photonics. Thank you to Bryan and Jasper, two phenomenal long-haired engineers.

An endless thanks to my friends Milad and Alanna and to their lovely families, thank you because being able to say that I have a family on the other side of the ocean was a gift I did not expect to receive. Thanks to Salar for his loud laughs.

Thank you to my two girls Alycen and Michelle, your sweet smiles will remain etched in my memory as well as our crazy dances at Inner Harbor and thank you to Edu, Derek and Josh without which the house at 3313 Chestnut Avenue would have never been such a beautiful place.

Thanks to Christos and Kate with whom I had the opportunity to share the

emotion of presenting this work at my first conference.

Thanks to Mary Ellen for the exquisite dinners and the endless delicacy.

I really want to thank who, even if out of the academic environment, has been, and still is, one of the best teacher I have ever had. Thank you Christopher for teaching me that spontaneity, respect and mutual trust can open you to the most beautiful side of the world. That area where labels do not exist, where the stories you share are worth more than the clothes you wear. Keep on rocking my friend, I miss your compulsive drinking pineapple fruit juices.

Thanks to those who have always been and will always be the solid base on which everything is built, my mother and my father. You started by giving me life and now you are constantly feeding my passions and making my dreams possible. Thank you mom for being the sweetest adult that life has ever given me the chance to meet. Thanks for all the patience when it was 8 pm, dinner was ready but as long as the last page of art history was not repeated nobody started to eat. You taught me perseverance, you taught me that alone you are never as happy as when you are in two. Your benevolence is something I really hope to own one day, I carry you in my heart and in that thick curly hair of which I complain so much but that it makes me very pride to look like you. Thank you dad for never having stopped to be my first supporter. You have never understood anything about dancing nevertheless I was expecting nothing more than your comments after each concert. Thank you for teaching me to be patient, to work silently, that the results come to everyone if you deserve them and that you do not ultimately have to give an account to anyone but yourself. I will never stop being your princess. Thank you because no one more than you both has taught me that you should never forget where you come from and that the ideals should never be thrown to the wind, that only by following yourself you may one day "be great".

Thank you Ale, for never saying "I love you", there is no need between brother and sister, we just know that we will always be there for each other. Thank you for being able to never have made me doubt this.

Thank you to all my extended family, the super ninety "nonna Gina" that is still able to give me some Latin grammar lesson, "nonna Rosi" that I miss so much along with all the garlic that she put in dishes that were so nice to share, my crazy aunt Manola I confess that every time Dad called me with your name I was proud to be exchanged with you, I love you. Thank to my cousin Jacopo, a mix of stubbornness and sweetness, impossible not to love. Thanks to my uncle Massimo, Aunt Roberta, Leo, aunt Anna, uncle Giuseppe, Stefano and Ottavia for never having denied me a hug and a loving smile.

Thanks to "Gnagno", my long life friends, despite the distance, each one of your thoughts, has always arrived at the right time. Sissa, Cami, Arci, Gple, Gambo, Ari, Leo, Filo, Pollo and Berfro I am looking forward to playing TABU with all of you, this coming Christmas. Alice and her determination, eternal example of life,

the sweet Sofia that has crossed the ocean to give me the most beautiful hug I could wish for and Erika, my soul sister, a strong woman with a giant heart that has managed to never let me feel the distance between us, you three beautiful girls have a special place in my heart.

Thanks to my "adorate", because you will be my favorite dancers forever, your vital energy is a fundamental fuel. Let's see each other more! A special thank you to who has been there since the Dalmatian-shaped and dragon-shaped dinghies, I love you Ila.

Thanks to the irreplaceable Chiara, Alessia and Greta. To you I do not really know what to say, after all those hours of study, chat, laughter and lots of wine I can only say that without even one of you three, all the beauty I have now here in Turin would not exist. You are the people I have chosen to have next to me in this magical city for this you are now indispensable. It's reassuring to know I can always count on you.

A big thank you to my sweet neighbor who has always been able to give me happiness, laughter and lots of good food, thanks Elenina.

Thanks to my bodyguards Alessandro, Davide and Leo who have always turned Sunday lunches at "via Tripoli 64" into fun family reunions. Thanks Davide for helping me with physics, thanks Ale for being the male version of the craziest part of me and thanks Leo because the alarm clock with Good Times Bad Times will remain an indelible memory from the early years in Turin.

Thanks to Luca for the technical and moral support, I will never forget anything you have done for me.

Tanks to my little roommate Antonino, for being always willing to listen me.

Thanks to Gianchi for always bringing me home and never in silence.

I own a big thank you to all the fantastic people of the "Studio danza Ieva", place full of passion, and madness. I wonder if without that place I would have never met my blue-eyed Sofi. I want to thank you, for always welcoming me at your home with a smile and for getting me involved in your madness.

Finally thanks to you Fabio, for the respectful eyes with which you have looked at me for the last two years, those eyes that have guided me through my fears, heard me in the sleepless nights, transmitted me the energy of rock music and conducted me to the ends of the world. I wish to myself that life never makes me feel that I miss you, it would be difficult to stand. Thanks to your beautiful family because they can make me forget the distance that separates me from mine. The serenity I feel in "via Puccini 3" is priceless.

I hope I have not forgotten anyone.
Now it's the big leagues and I need all of you.
Thank you, I love you all.

Elena

Contents

Chapter 1

Introduction

In 1997 IBM supercomputer defeated the world chess champion but it won playing an unequal weapons game. In a single second the computer could evaluate almost 200 million possible moves while his human rival could take into consideration 3 moves per second at most. Even today computers can not really compete with human brain especially in fields such as hearing, vision, pattern recognition and learning [3]. After all, nature took millions of years to solve problems which traditional engineering approaches struggle to solve and this is why scientists and engineers started to look at biological systems trying to emulate or even surpass their performances. For that reason, from the late 1980s a huge number of researches and studies has focused on imitation of neuro-biological architectures present in the nervous system, leading to the development of the concept of Neuromorphic engineering. Neuromorphic Engineering is an interdisciplinary field of study that involves biologists and physicians as well as mathematicians and engineers and whose main focus is to design analog, digital, mixed-mode analog/digital VLSI, and software systems that implement models of neural systems [43]. In order to accomplish this, it is first required to explore phenomena and theories that explain the process you want to model: neuromorphic algorithms generally lead to the elaboration of computational models that among others more practical applications, provide important insight into various mechanism in the brain.

One of the mechanisms that most characterizes human brain is the so-called *selective attention*. *Selective attention* is the process of reacting selectively to certain stimuli, when several occur simultaneously and it characterizes each of our senses. It is a passive and instinctive phenomena different from *mental concentration* in which personal will is involved. Evolution has made it possible because the brain must deal with a large amount of sensory input at each moment of our existence and it must decode them in real time without having enough resources to process all the information with the same degree of detail. In general human attention occurs after a quick scan of all the stimuli coming through our five senses from the external environment and then guides the sense in question towards the most salient stimulus.

The mechanism of attention allows us to be physiological functional and efficient in everyday life. For that reason understanding and being able to predict how and where human attention is focused are important research topics in neuromorphic engineering.

A sensory path that carries a great amount of information is sight: it has been estimated up to $100Mbps$ carried by each optic nerve [28]. Light enters in human eyes through the pupil and it is converged onto the retina by the cornea and the crystalline lens. The retina is placed in the back wall of the eye and its receptors detect light energy and transduce the action potentials that travel along the optic nerve. The fovea, center of the retina, has the highest resolution in the eye and processing the entire visual field with its level of detail would require greater computational resources than the ones available. *Visual selective attention* has been described in 1979 by Shulman [31] et al. with the similarity of the spotlight in a dark room: looking at a particular area can be seen as directing a spotlight (the fovea) to a particular zone of a dark room. Today it is known that this explanation refers to what is called *overt attention*, which occurs every time the observer shifts his/her gaze (focuses his/her attention) to a different area of the scene that he or she is looking at. *Overt attention* differs from *covert attention* that occurs when an area of the scene attracts our attention without looking directly at it (as when your name is listed in a series of written words). In any case it is believed that these two types of visual attention are not independent: it is not possible to direct *visual attention* in a particular area of the scene, looking elsewhere. Factors that drive human *visual selective attention* attention can mainly be divided in two categories [8].

- **Bottom-up factors** that only depend on the scene the observer is looking at and are the results of the fact that some stimuli are visually more stimulant then the others and attract attention involuntary [27];

- **Top-down factors** that are controlled by the organism itself and its internal state [27].

Even if first computational models that tried to copy human *visual selective attention* mechanism date back to the 1980s, until few years ago they mainly were theoretical studies. Only in the last 15 years computational power allowed to look at more practical applications for these models. Despite this short history, computational models for *visual attention* are so many that listing them all is hard as well as not very useful for the purposes of our work. All of them have the final purpose of predicting which are the most salient areas of a scene given as input to the algorithm as a picture that represents the scene itself. General structure is repeated almost unchanged in the most famous models and follows psychological theories as Feature Integration Theory (FIT), despite this overall similarity, a strong dependence on the choice of parameters causes small changes in computer code to originate different

models with different performances. The first common output of every computational models is a *saliency map*. The concept of *saliency map* is born in 1985 [19] and refers to a two-dimensional map where pixels values are proportional to how much the corresponding area attracts observer's visual attention. A turning point for the state of the art can be considered the work published by Russel et al.. As far as we know, it is the first example of a computational model biologically plausible that does not describe the *visual attention* as a process guided exclusively by the image features but also by the organization of the image in perceptual objects that are recognized without requiring to have access to all their possible features [28]: the so called proto-objects. The initial decomposition of the input image in a series of channels link to elementary features such as intensity, color opponency and orientation is left unchanged from most of the previous models but biological concepts of Border Ownership Cells and Grouping cells are introduced.

This current study proposes a new computational model of proto-object based visual attention completely developed in MATLAB 2017b environment. Its backbone is the same as the one suggested by Russel et al. but it aims to add a new channel (disparity channel) carrying depth perception information. Despite the large number of models of visual salience already mentioned, it is surprising how studies are largely focused on a two-dimensional representation of reality, primary features are generally chosen without taking into consideration that we live in a three-dimensional world and that relative arrangement of objects in space could drive our attention. One organizing principle that humans employ is to structure the world in parallel planes or surfaces. Each plane is associated with a particular distance from the observer. First experiments conducted on maquaque monkeys at the Brain Institute of the Johns Hopkins University seem to confirm the theory that the mechanism operated by the grouping cells can also be adapted to a three-dimensional analysis, this makes proto-object model valid for spatial analysis. If little research has been done on three-dimensional models of feature based saliency three-dimensional models of proto-object saliency are a almost new field of study. Differently from a previous work that assumed that a depth map is pre-computed, this model only assumes stereoscopic information as input. Two views of the same scene are given to the algorithm, they are identical except for the fact that the same image point is horizontally shifted between the two views, the shift entity is inversely proportional to the distance of that specific point from the observer. This reproduces what happens in humans between left and right eye. The model is built on the assumption that the pop-out effect characterize depth human perception as well as it happens with other elementary features: depth discontinuity attracts human attention more then regions lying at the same distance from the observer. A complete computational model usually has as final output a series of image regions that tries to reproduce human saccades and starts with the most salient area of the saliency map. Anyway our study ends with the saliency map obtained by the combination of two different

maps.

1. **Two-dimensional saliency map** that is obtain from one of the two view given as inputs and includes contributes of intensity, color opponency and orientation channels.

2. **Three-dimensional saliency map** that is obtain from both views given as inputs and includes contribute of disparity channel.

The algorithm works by searching for disparities between the two stereo images but disparity values change with changing in images resolution, making model automatization difficult. For that reason the disparity channel has been first tested by using artificially created images where disparity values were known. Resulting saliency maps show the ability of recognizing depth discontinuities (associated with disparity values rarer than others), regardless their position and frequency with which they occur. In order to test any improvements made by introducing disparity channel in the computational model of proto-object based saliency a dataset with both stereoscopic views and a valid ground truth (i.e.fixation density map estimated using an eye-tracker) is needed. Only one dataset composed by 18 couples of images, was available online. Even if all the three metrics used to evaluate the model show better or at least equal ability of predicting a saliency map that is as much as possible similar to the map provided as ground truth, the paucity of datasets does not permit to these results to reach statistic significance. Once a new series of datasets will have been made, the model aims to automate the choice of disparity values to look for. It will follow the working method of far, near and tuned-zero cells of the visual cortex that respectively respond to disparity in planes further away from the fixation plane, closer than the fixation plane and on the fixation plane.

Dealing with a large number of data that has to be collected and analyzed is not an exclusive problem of the human brain: as widely known it affects many modern technical systems. Computer vision systems are asked to deal with a number of pixel values that can reach a few millions per frame and so computational cost is generally very high and task becomes particularly difficult if real time application is needed such as in cognitive systems and mobile robotics. If we think about a robot that is able to autonomously drive a car having a limit amount of resources and facing an unknown environment, the ability to prioritize inputs coming from the outside is of fundamental importance for the reduction of the computational cost and complexity. Moreover, with the massive and domestic use of both computers and Internet, predicting how and where visual attention is directed when stimulated by an image on a screen, could revolutionize the advertising graphics. Furthermore, psychology of colors applied to the brand as well as the choice of particular fonts for communication or Gestalt psychology [21] are notions already widely diffused in this field.

Chapter 2

Background

Thinking about how the brain works, human beings receive a long series of multisensory inputs, such as what is seen, touched or heard, in a completely indirect way. All these sensations are experienced without thinking: the brain uses much less power to function than a current microchip. As already anticipated in chapter 1 this is possible thank to a brain function that allows to instinctively select only some of all the environmental stimuli. This mechanism is called *selective attention* and it is typical of every human sense, including the sight that is the focus of this study. In order to understand and imitate *visual selective attention* biologist, doctors and engineers are working together to elaborate a computational model of visual saliency that is both functional and biological plausible. Looking for the definition of the word "saliency" (or "salience") in the dictionary it is described as "the quality of being particularly noticeable or important". In fact, the purpose of the computational models just mentioned is precisely that of predicting and simulating how and where visual attention is directed. In this sense, the recent development of the state of the art has led to the emergence of a new class of neuromorphic algorithms for the simulation of *visual selective attention*. Starting from studies on biology, physiology and human psychology these models look at *visual saliency* as an object based process instead of a feature based one (see section 2.3 for further information) and a sufficiently in-depth knowledge of the mechanism at the two-dimensional level has moved the focus of researches to computational models of *visual saliency* that look to the three-dimensional space.

2.1 Human Visual Attention

Human beings are constantly bombarded by a large amount of sensory stimuli including the ones coming from the retina, fundamental component of human vision system whose cells transform light energy in electric potential. Therefore, also the visual system is subjected to the mechanism of attention operated by the brain,

which in this case has the specific name of *visual selective attention*. *Visual selective attention* is set when competition between stimuli coming from the visual field is solved in favor of one or few elements that are elected as carriers of more relevant or more noticeable information [6]. *Attention* determines the order in which an image or a scene is investigated. Even though our overall visual memory is considerably long and good, the reaction to small changes in the visual field, including an image on a screen, is poor. This statement is supported by experiments on change blindness conducted since the nineteenth century. If the image is complex, only the general sense is retained and at each moment only a small part is analyzed while the rest is ignored. This part is often, but not always, the region you are looking at. Background and small details are generally ignored. Anyway thank to *selective attention* people automatically elect salient regions (points of main interest) in their surrounding and explore scene by rapidly change their focus.

There is not only one specific area of the brain involved in attention process: one of the most relevant outcome of neuro-physiology on *selective visual attention* is that there is not an unique area of the brain that drives attention mechanism. Neurons correlated to visual search appear to be spread in all brain areas link to visual processing. Rather, this mechanism involves a network of anatomical areas. What is known for sure is that information travels in a parallel way: different features are processed by different zones of the visual cortex. The exact processing of information within the visual paths is not known yet but studies in this field are proceeding. Most common belief is that there are three main pathways: color, shape and motion which is also responsible for depth processing. Anyway lateral connections between different areas of the visual system show that pathways are not completely separated.

From the entrance of the light stimulus trough the *retina*, back wall of the eye, up to the higher brain areas where the signal is interpreted, the region of each anatomical area is projected on the subsequent area as if the two were superimposed. The signal travels along the *optic nerve* whose fibers not always are neatly collected but each time they arrive in a new anatomic area of the brain recompose to always end in an equally orderly manner. In order to elaborate a computational model that simulates the mechanism of *visual attention* is fundamental to have at least a brief knowledge of the human visual system and pathways [11].

2.1.1 Human Visual System

External sensory organ of the visual system is the eye. It collects light from the external environment and through light receives information. Light enters through the *pupil* and its intensity is regulated by a diaphragm whose name is *iris* and thank to a series of lens, light signal is focused on the *retina* to form the image. For an overall view of human eye structure and the relative position of its main components, refer to the figure 2.1.

The Retina and the Optic Chiasm

The *retina* transforms light inputs into electric signals that are sent to the brain by the *optic nerve* so it is not a peripheral structure but part of the central nervous system. At brain level signals are first elaborated and then interpreted. The *retina* is situated in the back of the eye and is sensitive to the light thank to the over 100 million photoreceptor cells, *rods* and *cones* [11].

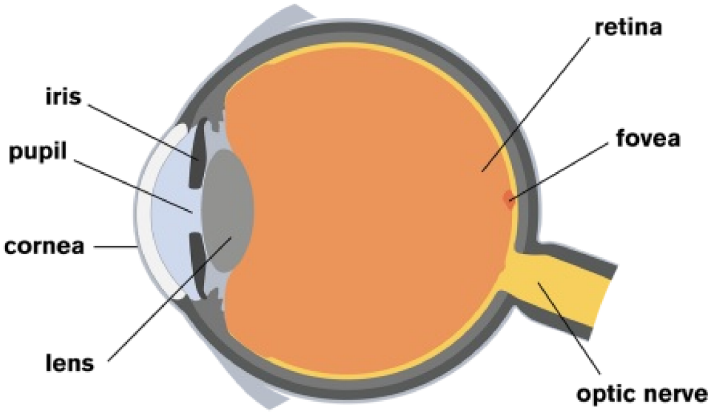


Figure 2.1: Illustration of human eye structure: black bold writings refer to its main anatomic components name; black segments indicate the position of each component [34].

Rods, photoreceptor cells in the *retina* are:

- more numerous;
- more sensitive to light: suitable for night vision;
- not sensitive to color.

Cones are less sensitive to light than *rods* so they are suitable for daytime vision, they provide color sensitivity and among them there are three different types of color reception.

- **S-cones or Short-wavelength cones:** maximum absorption at wavelengths between $400nm$ and $500nm$ (blue portion of the visible spectrum);
- **M-cones or Middle-wavelength cones:** sensitive to wavelengths between $500nm$ and $600nm$ (green portion of the visible spectrum);
- **L-cones or Long-wavelength cones:** sensitive to wavelengths greater than $600nm$ (red portion of the visible spectrum).

In the central part of the *retina* there is a small pit with a circular shape and about $1.5mm$ in diameter: the *fovea*. Here there is the maximum concentration of the *cones* (but only those that allow the vision of red and green), while the *rods* are completely absent. The *fovea* is the region with the maximum visual sharpness. Because of the distribution of photoreceptor cells, we see only the small region currently fixed in a high resolution, the whole surrounding is representing with a very law precision. In order to amplify the total area perceived at high resolution, light stimuli from different areas of the visual field, arrive at the *fovea* subsequently. This is mainly possible thank to eye movements. Inter-neurons bring information from photoreceptor cells to *ganglion cells* that operating on the chromatic inputs determine color and luminance opponency [11]. Defining the *receptive field* of a sensory neuron as a small area in the sensory space (e.g., the body surface, or the visual field) in which stimulus modifies neuron firing, *ganglion cells* in the *retina* has a circular *receptive field* separated in two concentric areas: center (internal area) and surround (external area). There are two types of *ganglion cells*.

- **On-center ganglion cells:** massive reaction (number of action potentials per second) for a light stimulus located in the center of their receptive field and inhibition caused by a light stimulus located in the surrounding area, suitable to well see a bright region on a dark background;

- **Off-center ganglion cells:** massive reaction (number of action potentials per second) for a light stimulus located in the surround of their receptive field and inhibition caused by a light stimulus located in the central area, suitable to well see a dark region on a bright background.

These two types of *ganglion cell* are equally present in the *retina*. In figure 2.2 the mechanism of *on-center cells* and *off-center cells* reaction to two opposite stimuli (bright center and bright surround), is briefly summarized.Some of this cells are also sensitive to color contrasts: blue-yellow and red-green.This *receptive field* organization can be computationally modeled by a difference of gaussian filters (this concept is further debated in chapter 3). After passing through the *retina* the visual information, driven by the optic nerve (continuation of the ganglion cell axons), runs to the *optic chiasm*. In this area of the brain a partial crossing between the nerve fibers, takes place. From there two pathways go to each brain hemisphere (figure 2.4):

Cortical or Retino-geniculate pathway that passes through the *Lateral Geniculate Nucleus* (LGN) arrives to the *primary visual cortex* (V1), at the back of the brain and carries 90% of the global visual information.

Subcortical or Collicular pathway that does not cross the *primary visual cortex* (V1) but passes through the *Superior Colliculus* (SC).

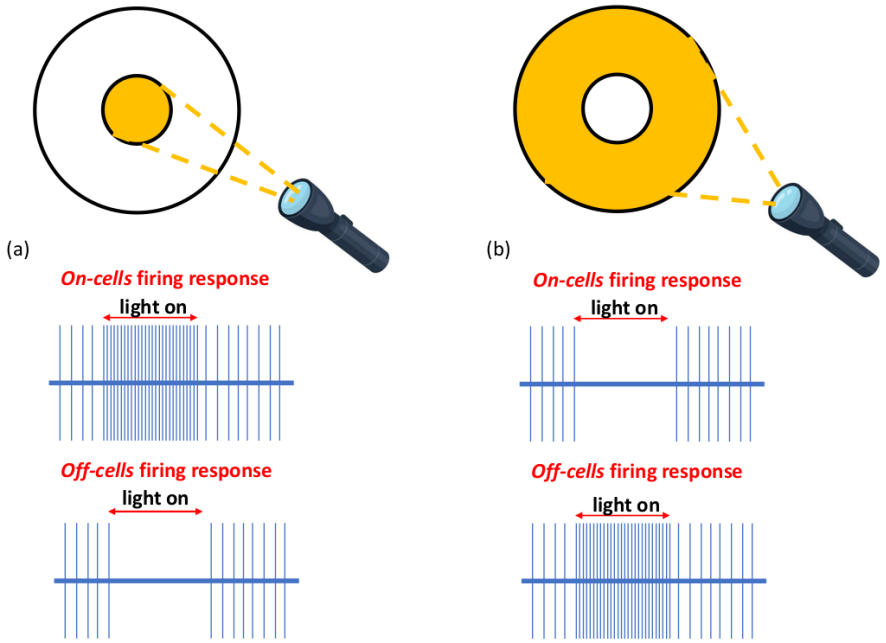


Figure 2.2: Briefly overview of ganglion cells reaction to different stimuli in their receptive field: flashlights and light beams indicate the illuminated area og the receptive fields (yellow areas).(**a**) The bright zone of the receptive field is the central area: on-center cells are stimulated while off-center cells are inhibited. (**b**) The bright zone of the receptive field is the surrounding area: on-center cells are inhibited wile off.Figure adapted from [20].

Cortical or Retino-geniculate Pathway

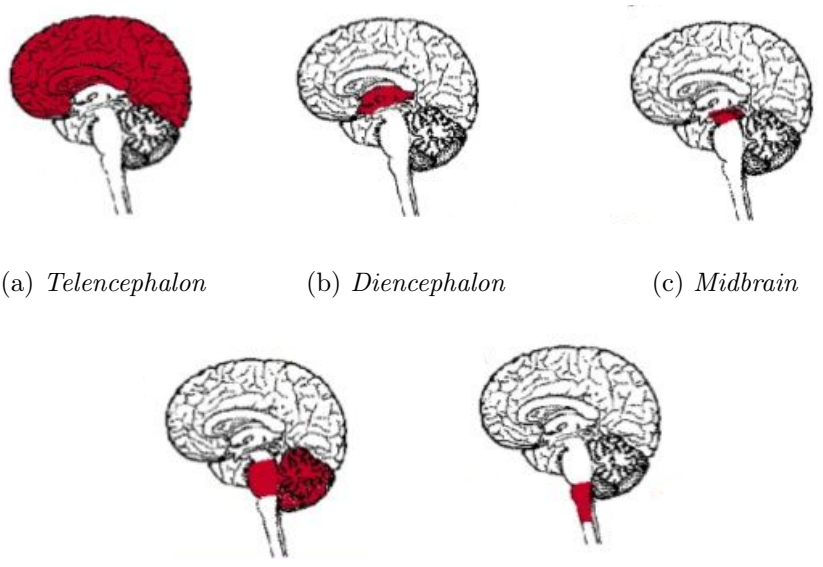
Lateral Geniculate Nucleus is composed by six cellular laminae and its cells have a receptive field with the same shape as the one of the retinal *ganglion cells* but larger and with a stronger surround [11] [10]. From *Lateral Geniculate Nucleus* the visual information is directly transmitted to the *primary visual cortex* (V1) that is the access way to the brain and the most investigated area in visual system. In V1 there are three different types of cells.

- **Simple cells**: orientation sensitive;
- **Complex cells**: take inputs from the simple cells, larger receptive field than the simple cells, sensitive to moving lines or edges;
- **Hypercomplex cells**: take inputs from complex cells, capable to detective lines of a particular length or line that end in a specific area.

Up to the primary visual cortex the processing stream is called *primary visual pathway*.As information travels along the two paths, it crosses more and more specialized cell populations able to elaborate the numerous outputs coming from the previous cells.

Subcortical or Collicular Pathway

Visual inputs from the *retina* can directly arrive in the *Superior Colliculus* (SC) structure. The SC is located in mammalian midbrain (figure 2.3): its superficial layers (sSC) receive visual signals while the deeper layers (dSC) are active in the eye movement orientation. In sSC neurons do not show any particular predispositions for stimuli linked to specific features, some of them show to motion without any preference on directions. After the SC signals arrive in the *pulvinar* (Pulv.), relief of the posterior end of the diencephalon (figure 2.3) that whose neurons are similar to the ones in sSC. From the *pulvinar* area the subcortical pathway reach the *Frontal Eye Field* (FEF) that is a region in the frontal cortex with a robust eye movement related activity [1].The *FEF* sends signal back to the deeper layers of the *Superior Colliculus* which communicates directly with the brain-stem (situated between hindbrain and the mylencephalon, see figure 2.3). Eyes move, under the command of the brain-stem.



(a) *Telencephalon* (b) *Diencephalon* (c) *Midbrain*

(d) *Hindbrain* (e) *Mylencephalon*

Figure 2.3: Subdivision and organization of the human brain: visual pathways groups of cells are located in different areas of the brain (**a**) Telencephalon area highlighted in red. (**b**) Diencephalon area (where Pulv. is located) highlighted in red. (**c**) Midbrain area (where SC is located) highlighted in red (**d**) Hindbrain area highlighted in red. (**e**) Mylencephalon area highlighted in red.

Extrastriate Cortex

From the *primary visual cortex* information is sent to higher brain areas globally called *extrastriate cortex*, in order to differentiate them from the *primary visual cortex* that has a striated architecture. Optic signals travel in the *extrastriate cortex* following two different pathways (figure 2.4):

"What" Pathway or Ventral Stream that is the color and form pathway. It goes through all the different areas listed below, reaching the Infero-teporal cortex where objects recognition takes place (it carries the *"what"* information). Signals cross brain regions in the following order:

1. **V2**;
2. **V3**;
3. **V4**;
4. **Infero-temporal cortex (IT)**.

"Where" Pathway or Dorsal Stream that is the motion and depth pathway (it carries the *"where"* information, processed by the Poster-Parietal cortex). It goes through all the different areas listed below, according to the following order: order:

1. **V2**;
2. **V3**;
3. **V5 or Middle Temporal area (MT)**;
4. **Parieto-Occipital area (PO)**;
5. **Poster-Parietal cortex (PP)**.

Visual pathways have been described starting from the eye and proceeding up to the higher regions of the brain involved visual stimuli interpretation, but it reality the overall mechanism is bidirectional: top-down connections go from the higher areas to the LGN. Also different pathways are not so strictly separated as well: lateral connections make V4 ("What" pathway) and MT ("Where" pathway) communicate.

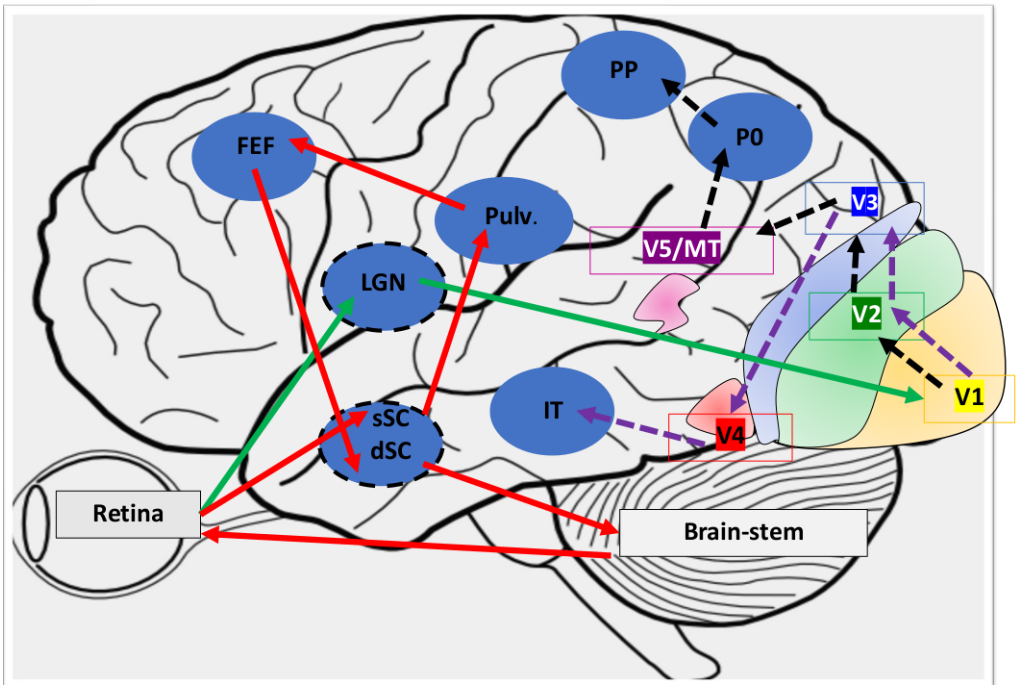


Figure 2.4: Visual pathways overview: continuous arrows mark primary visual pathways (up to V1), green arrows follow the *Cortical Pathway*, red arrow follows the *Subcortical Pathway*; dotted arrows mark the *Extrastriate Pathways* (higher than V1), purple arrow follows the *"What" pathway*, black arrow follows the *"Where" pathway* (sSC and dSC: superficial and deeper *Superior Colliculus*, LGN: *Lateral Geniculate Nucleus*, PO: *Parieto-Occipital Cortex*, PP: *Parieto-Parietal Cortex*,Pulv.: *Pulvinar*, FEF: *Frontal Eye Field*, IT: *Infero-Temporal Area*, V1-MT: from *Primary Visual Cortex* up to *Middle Temporal cortex*).

2.1.2 Neurobiological Correlates of Visual Attention

As anticipated in previous section, how and which brain areas are specifically involved in various steps of *visual selective attention* is still an open biomedical research question. One of the most important conclusions was the understanding that almost all the areas involved in visual processing take part in this mechanism. In order to simplify the search for an answer to this open question in the field of research on perception, some studies divided attention process into three consecutive steps or functions [24]:

1. orienting attention;
2. target detection;
3. alertness.

First function involves three brain areas: the *Parieto-Parietal cortex* (PP), the *Superior Colliculus* in the midbrain and the *pulvinar* in the diencephalon (see figure 2.3 and 2.4 for reference). The *Parieto-Parietal cortex* takes care of diverting attention from the previous focus location (inhibition of return), the *Superior Colliculus* moves the attention to a new location and the *pulvinar* reads data regarding the new attention location. This combination of systems is addressed as *posterior attention system*. What is called by Posner and Peterson in their work *anterior attentional system* deals with target detection function [24] and they support that the brain frontal areas are involved in this task. Finally high priority signals alertness depends on brain activities in the midbrain zone (see figure 2.3). Eye movements is guided by *Superior Colliculus* and *Frontal Eye Field* (figure 2.4). Later studies have also focused on understanding which areas of the brain are stimulated by top-down rather than bottom-up factors (see section 2.1.3) and it has once again emerged that there is no single area involved but a network of areas. This network includes areas of the brain that are verified to present activities in visual search tasks.

2.1.3 Bottom-up vs Top-down Attention

The mechanism of visual attention as well as that of simple perception is driven essentially by two processes and the related factors.

- **Bottom-up process and factors**: attention is guided by the emerging characteristics of sensory information.
- **Top-down process and factors**: attention is underpinned by intentional and conscious processes.

Some studies following this distinction, divide attention itself in two type: *Bottom-up* and *Top-down* attention. *Bottom-up* factors are derived exclusively from the visual perceptual scene. In this sense when we refer to "salient" areas, we generally refer to areas of the scene and / or of the image we are looking at, which in an entirely involuntary way attract our attention. This involuntary process starts from the sensory organ (generally called the "lower" level of the entire system involved in the attention mechanism), eye in visual system, and proceeds up to the brain areas where signal interpretation takes place ("higher" level of the system). This overall path that goes from the bottom to higher levels, gives the name (bottom-up) to the whole process. Common examples are a high-contrast area against a background or uniqueness of a specific region within the scene. A bottom up factor can not be voluntary neglected, this effect is called *attentional capture* and it is a very important effect that can save our life if an emergency bell starts ringing capturing our attention or a fire stars burning [10]. On the other hand *top-down* attention refers to a process that starts from cognitive factors as preknowledge, context, expectation and current goals [11]. This mechanism is sometimes distinguished from its bottom-up

antagonist because it is described voluntary rather than automatic. In my opinion this statement, although partially true, can be misleading. In fact, unless we are asked to look for a figure, a color or something specific within the scene in front of our eyes (in other word we are asked to perform a specific task), even top-down attention is not a mechanism that depends strictly on our will. For example, if we are hungry in a room where something edible is present, we will give visual attention to food, but this, although guided by our brain, does not really depend on our active will. *Visual search* is the best known and investigated aspect of *top-down* attention. We experience it every time we have to look for a friend in a crowd. *Bottom-up* process has been more studied than the *top-down* one and even the model shown in this study is focused on *bottom-up* attention, this is due to the fact that data (such as image features) driven process are easier to control than cognitive factors (such as knowledge and expectations) [10].

2.1.4 Visual Search and Pop-out Effect

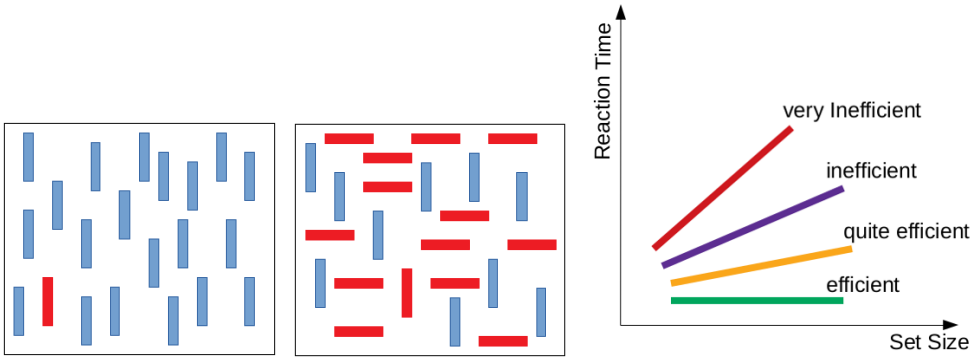
Visual search is an experimental paradigm consisting of presenting a set of objects in the middle of which, in half of the tests, a target object appears. The observer task is to report whether the target is present or absent. In a variant of this paradigm, the target is always present but instead of having a single defining attribute, it assumes one of two possible in each single test. In this case, the task of the subject is to report which attribute of the target has been presented. *Visual search* can be briefly summarized as the search for a target object (having a specific required feature) positioned between a series of distractors (objects that do not have the specific feature that the subject is asked to find). The efficiency of *visual search* is measured from:

Reaction Time (RT) needed to find the target among a certain number of distractors, amount of time in between the presentation of a stimulus and the issue of an answer;

Accuracy (AC) that is linked to the error rate achieved in completing the experiment.

Efficiency can be estimated from the slop of the curve that represents the *reaction time* of any single search, depending on the dimension of the set on which *visual search* is operated. Obviously, since the target is generally only one object, enlarging or reducing the set means increasing or decreasing the number of distractors, therefore making the search respectively more difficult or more easy.However, *visual search* experiment (until research is still possible) a physiologically healthy *reaction time* does not exceed $2000ms$. In figure 2.5 two examples of *visual search* experiment are shown. If the target differs from the distractors, only because a single feature (figure 2.5 (a)), *visual search* becomes a *feature search*, if it differs for more

than one feature (figure 2.5 (b)) it is called *conjunction search* and is less efficient. *Feature search* occurs and in parallel across the visual field while *conjunction search* occurs serially and requires attention [27]. When the target is well defined by a single characteristic, ex. a red bar between blue bars, the *"pop-out"* effect occurs (figure 2.5(a)). If this happens TR does not vary with the increase in the number of distractors. Efficient visual searches are often linked to this phenomenon. Common sense suggests that the more an object is similar to the background, the more difficult it is to identify it through the process of selective visual attention (unless particularly relevant top-down factors take place). For that reason, computational models that aim to find salient zones in any scene (or image), must evaluate how much for each elementary features of the image, *visual search* can be approximated to the *"pop-out"* effect.



(a) *Feature search* (b) *Conjunction search* (c) *The continuum search slopes*

Figure 2.5: **(a)** Feature search: the target (red bar) differs from the distractors (blue bars) by one feature (pop-out effect). **(b)** Conjunction search: the target (red vertical bar) differs from the distractors by two conjunct features. **(c)** Reaction time of a visual search experiment in a function af the set size, curve slop indicates the efficiency (figure adapted from [10]).

Reaction time does not give exhaustive information about the search process itself. Measuring *accuracy* and seeing how it varies changing the *stimulus onset asynchrony* is another way to evaluate search efficiency. *stimulus onset asynchrony* (SOA) is the time in between two stimulus, considering the first one as the actual search stimulus and the second as the mask that terminates the search. If the search is easy, short SOAs do not preclude an efficient experiment success, more difficult searches require longer SOAs.

An important field of study concerns the so-called *search asymmetries*. It is now well established that looking for an objective target A in the midst of distractors B is not always equal to the search for target B between distractors A. In fact if between configuration A and configuration B, one is the canonical situation (most common) then it will be easier finding an object that distinguishes itself for a rare

peculiarity rather than finding an one with a common aspect among distractors with a peculiar appearance. Searching for a tilt bar (rare situation) among strictly vertical bars (common configuration) is harder than the inverse search this is also due to the fact that detecting the presence of a characteristic is easier than detecting its absence (figure 2.6). Furthermore due to the *retina* configuration and the central position of the *fovea*, the highest resolution area of the *retina* (see section 2.1.1), makes target at peripheral locations more difficult to detect so both reaction time and errors increase with targets distance from the center [10] (*eccentricity effect*).

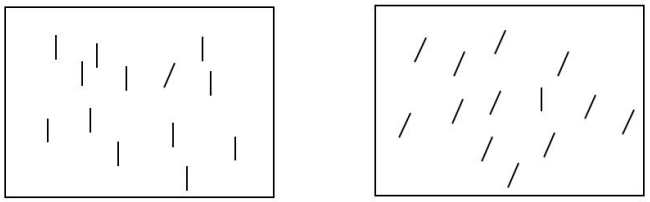


Figure 2.6: Example of different configuration for a visual search task: left image shows the situation of an uncommon target among common configuration distractors; in right image the situation is the opposite: common object is the target to find among peculiar distractors. The observer should experience more easy the visual search in the left panel.

2.2 Human Depth Perception

As already mentioned in subsection 2.1.1 visual signal elaboration starts when light coming from the external environment, is focused on the *retina* from the lens system of the human eye. At the *retina* level a two-dimensional image is created. However, common experience suggests that the perception we have through the visual system is three-dimensional as the external word itself. Perceiving depth as well as the ability to see the world in three dimensions means being able to evaluate external objects distances. In order to reconstruct three-dimensional aspects of the environment, visual system uses a series of depth cues: information from the surrounding. Depth cues include monocular and binocular depth cues.

- **Monocular depth cues:** perceivable using just one eye;
- **Binocular depth cues:** perceivable only using both right and left eye.

Binocular cues contribute to have a more accurate view of the three-dimensional space while the monocular ones are rather artifices thanks to which our visual system allows us to perceive depth even using only one single eye. This results from the elaboration that our mind makes of data recorded by visual organs and their links with previous sensory experiences and acquired cognition memory. A list accompanied by a brief description can help to understand this statement, the usefulness of depth cues and the difference between the two classes.

Monocular Depth Cues

The information coming from each of the two eyes is sufficient to have a depth and distance estimate. The best way to understand how the brain is able to give us depth perception without using two eyes and stereo-vision (see subsection 2.2.1) is to take a look at what are the monocular cues and their mechanisms. The different molecular cues are listed below [39].

- **Motion parallax:** most used by some animals rather than by humans (e.g. some species of birds). Generally when an observer moves can obtain information about external objects distance by observing their relative motion against a fixed background: closer objects zoom by slower than the more distant ones. This is clearly perceived while traveling by car: pedestrians (close to the car itself) appear to move really fast while threes (that generally are distant from the road) seem to move slowly.

- **Depth from motion:** if an external object is moving toward the observe, its distance as well as its motion is estimated observing how its size changes. This happens because retinal projection of the observed external object expands while it moves.

- **Kinetic depth effect:** when a three-dimensional object is in front a source of light and its shadows is reflected on a screen and we see it from the other side of the screen we see it as a two-dimensional shape but once it rotates visual system has the necessary information to perceive its third dimensions.

- **Perspective:** lines that run parallel seem converging in the distance allowing us understand, in a more or less detail way, the relative distance of two part of an object or of features in a landscape. This is clearly visible when we look at a long road that goes straight in front of us and that seems becomes more narrow as long as it goes off in the distance.

- **Relative size:** only knowing two objects have the same size and without having any additional information (such as the effective size) realizing that one appears bigger that the other let the observer comprehend that it is closer.

- **Familiar size:** the fact that visual angle projected onto the retina decreases with the distance, combined with previous knowledge of the dimensions of a particular object, let the observer determine the absolute distance of that particular object.

- **Absolute size:** even if there is only one object in the scene and its size is unknown, a smaller object seems more distant than a bigger one that is presented at the same location.

- **Aerial perspective:** due to the presence of atmosphere that causes light scattering, objects at a great distance have lower luminance contrast and lower color saturation so they appear blurry.
- **Accommodation:** for distances up to $2meters$ when we try to focus on a far away objects some intraocular muscles stretch the eye lens changing focal length. The signal of muscles moving in sent to the *visual cortex* that use it to evaluate depth/distance.
- **Occulation:** when near surfaces overlap far ones observer perceive them as closer. This cue leads to a nearness "ranking" .
- **Curvilinear perspective:** the fact that at outer extremes of the visual field parallel lines curve, is usually eliminated both in pictures and in paintings but it really makes the observer feel like he or she is in a three-dimensional space.
- **Texture gradient:** this monocular cue is one of the most experienced in everyday life. To explain it, it is enough to make the example of a flowers field: when seen closely the individual components are clearly visible while when viewed from a distance the individual flowers are no longer distinguishable and rather the field looks like a single mantle (or spotted homogeneously if there are flowers of several colors), fine details are visible from closely.
- **Lighting and shading:** how light falls on object helps our brain to understand how far from the light source the object is and to determine its shape.
- **Defocus blur:** depth focus of the human eye is limited so blurring of distance objects helps the brain to estimate distance and depth even if all the other cues are not available.
- **Elevation:** if the object is visible together with the horizon we perceive it closer to us if it is farther from the horizon or farther from us if it is closer to the horizon. Moreover, if the object is near the horizon and it moves up (in a position higher than the horizon) or down (in a position lower than the horizon) it seems to move closer to the observer who is looking at it.

When the two eyes do not work together for the visualization of the same image, depth perception is limited and less precise but visual field is bigger. In fact monocular visual fields of the two eyes partially overlap in the binocular area. However, a healthy human being is able not only to exploit all monocular cues and combine them to estimate distances and three-dimensional space, but also to integrate them with binocular vision for greater precision in depth transmission. So few but more precise binocular cues allow humans to better perceive depth and third dimension. Visual acuity of the binocular vision is much greater than the monocular one, it could reach more than double and up to about 240% as maximum value.

Binocular Cues

Human eyes are about 6cm away one from the other and this makes their visual fields overlap, giving rise to some retinal correspondences that lead to the fusion of the two relative projections. When an observer is looking at a generic object, ocular axes converge in a point called *fixation point*. An image from every point of the object we are looking at, is projected on the *retinas* in a couple of points (one for each eye) called *retinal points*. A point located in the right part of the visual binocular field, has its projected image on the part of the right *retina* close to the nose and on the part of the left *retina* close to the temple. It is perceived as one single point located at the right. For that reason, at this first step of the visual pathway, observer already has information about object position in the visual field. The image that originates in the two eyes is perceive as one because for each eye, in the *retina*, there is a point having the same spatial value of an other point located in the contra-lateral eye *retina*. These points in both *retinas*, coupled by common visual direction, are called *retinal correspondent points* and are fundamental for the *fusion mechanism* [2]. This mechanism is composed by the following two types of fusion.

- Sensory fusion: psychological cerebral process that allows the unification of two similar images, of a fixed object, that originate at *retinal corespondent points*.

- Motor fusion: it contributes to keep the images at the *fovea* location thank to the alignment of the ocular axes (extrinsic musculature).

Corespondent points are not symmetric, they are coupled only by the *fusion mechanism*. All the objects that, at a certain time, are focused on the *retina* are all at the same distance from the *retina* itself and are all located on a imaginary curve called *horopter* (figure 2.7), all the points close to this curve (in front or behind) compose the *Panum's area*. This area is narrower near the fixation point and wider the more it goes towards the periphery. Obviously right eye sees more the right side of an object situated in the *Panum's area* while the left eye sees more the left side, so the images are not identical and neither have symmetric locations on the *retinas*. The *stereopsis* results from this small disparity between left and right view of the same object located in the *Panum's area*. Outside the *Panum's area* object-point is seen double because its image originates in the *retinas* in two areas that are not corespondent. Positioning the two index fingers in front of the eyes, one behind the other at a certain distance, is an easy way to demonstrate it. If we focus on the closest, the farthest appears double but if we focus on the farthest the closest appears double.

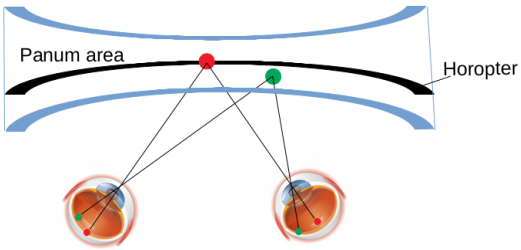


Figure 2.7: Horopter variation following fixation distance and Panum's area.

For what concern binocular cues, there are three of them that provide depth information while looking using two eyes. These cues are the following.

- **Stereopsis:** allowed by frontal separate locations of the eyes that causes they to have different points of view. Therefore, at any moment, there are two different retinal images of the same object available. Among other differences, a lateral displacement exists between the representations of a scene from the two views (left and right). From the comparison between these two images it is possible to obtain a very accurate depth perception. This concept is better explained in next section 2.2.1.

- **Convergence:** the two eyes located in different positions have to converge in order to look at the same location. This causes stretching of the extra-ocular muscles, kinesthetic signal helps in perceiving depth and/or distance. This cue is valid for distances up to $10m$.

- **Shadow stereopsis:** even if there is not actual stereoscopic parallax disparity between the two retinal images, a difference in the shading can cause the two images to be perceived equally as two stereoscopic images. This helps in depth perception.

Considering both *monocular* and *binocular* cues only *accomodation*, *convergence* and *familiar size* can give information about the absolute distance of an object, all the other cues are relative.

2.2.1 Stereopsis

Stereopsis means the three-dimensional vision that originates from the simultaneously stimulation of retinal horizontally different elements in *Panum's area* (figure 2.7). The minimal differences between the retinal images of objects placed in different planes in the *Panum's area* are used to capture the stereoscopic depth. Vertical disparity from the two retinal images does not cause any stereoscopic effect. This is the most complex expression of binocular cooperation and it is fundamental in order

to allow the subject to interact with the external environment. Distance and angulation with which object is fixated (for distances up to $30m$ [2]), are not perfectly equal: in normal physiological conditions the image originated from object fixation is projected in the *fovea* thanks to the convergence movements. Due to the fact that the distance between the two eyes is $6cm$ every object farther or closer than the *fixation point* projects its image at a certain distance from the *fovea*; nearest objects project their images on farther points of the *retinas* along horizontal direction and vice versa (figure 2.8). The distance between the *fixation point* image and other point image is called *retinal disparity*; visual system is able to evaluate this disparity and so give a sense of greater or less depth to the objects in the visual field.Stereoscopic perception seems to start suddenly when we are in between 3 and 4 months old, before in women than in men.

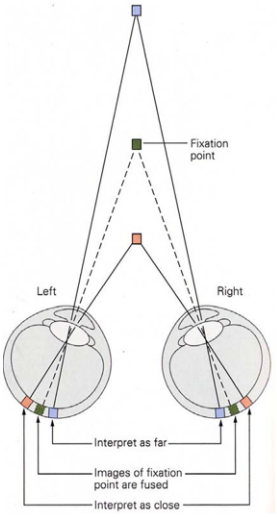


Figure 2.8: Retinal projection of the images from the fixation point, a closest point and a farther point [22].

Neurophysiology of retinal Disparity

Binocular neurons has been found both in primary visual cortex (*V1*) and in *extrastriate visual cortex* (especially in *V2*), these cells encode the degree of disparity between information coming from the two eyes. Some of these neurons respond selectively when the *retinal disparity* is caused by objects located in the *Panum's area* in a position closer to the observer than the fixation point, while others respond when the *retinal disparity* is caused by objects in the *Panum's area* in a position farther to the observer than the fixation point. Other binocular cells respond instead when objects in the visual field are at the same distance of the fixation point, but moved along the horizontal plane [22]. These three different types of *binocular cells*

are respectively called *near cells*, *far cells* and *tuned zero cells* [38] (figure 2.9 and figure 2.12).

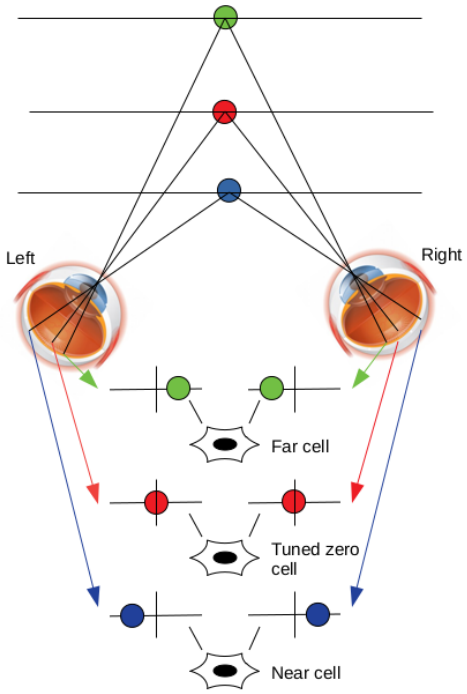


Figure 2.9: Schematic explanatory representation of the neurophysiological mechanism of retinal disparity interpretation operated by binocular neurons.

2.2.2 Space Organization in Parallel Planes

The *fixation mechanism* can be defined as the maintaining of the visual gaze on a specific point in space. When the ocular axes converge in a specific point this is called *fixation point*. As anticipated in section 2.2, this point lies on a plane parallel to the coronal plane of the head (figure 2.10) and that in three-dimensional space corresponds to the plane in which the two eyes are focused. Visual research experiments conducted on primates at the Mind and Brain Institute of the Johns Hopkins University (Baltimore, MD, USA), have shown that the mechanism of stereopsis in living beings endowed with binocular vision, including human beings, works by dividing the *Panum's area* of fusion into a series of planes parallel to the fixation plane located closer or farther to the individual himself, respectively (in front or behind the *fixation plane*). Even previous studies had already shown that one organizing principle which humans employ to understand how objects are related to each other and to themselves in space, is to structure the world in planes/ surfaces [12] [46]. Searching for a target within a plane of coplanar elements is efficient, see

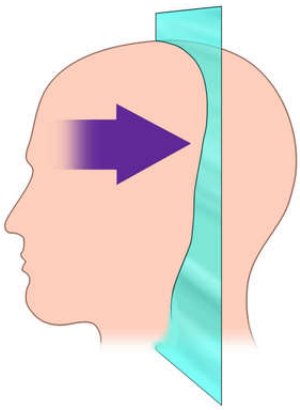
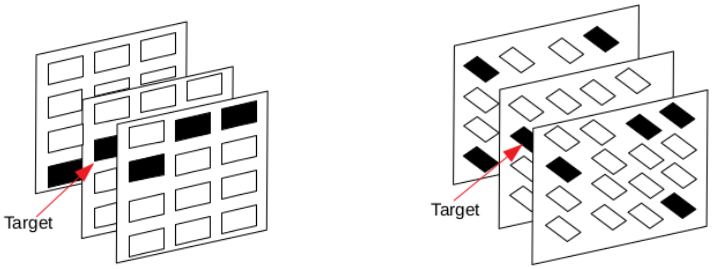


Figure 2.10: Position and orientation of the coronal plane of the head [14].

section 2.1.4 for information about *visual search* ad relative efficiency notion, while it is inefficient if the search can not be constrained to a surface (examples of different visual search tasks shown in figure 2.11 experimental data available in [46]). Furthermore, attention spreads across elements in a plane.



(a) *Easy Visual Search* (b) *Difficult Visual Search*

Figure 2.11: **(a)** Efficient Visual Search in the middle vertical plane: all the elements in the middle plane, including the target are coplanar with respect the search surface . **(b)** Inefficient Visual Search in the middle vertical plane: target slated more backward than the search surface [46].

For all these reasons the *Panum's area* should be redefined more as a part of the three-dimensional space, assuming therefore that this is a volume rather than an area. For the sake of simplicity, consider that this *Panum's volume* has the appearance of a parallelepiped having two of the lateral faces parallel to the fixing plane. These faces delimit the entire volume so that all the visual field between them is in the binocular fusion zone. The left and right retinal images from all the points included in this volume are seen as single. Three-dimensional scene (where fusion between the two retinal projections takes place) can be decomposed in a series of parallel planes as visible in figure 2.12.

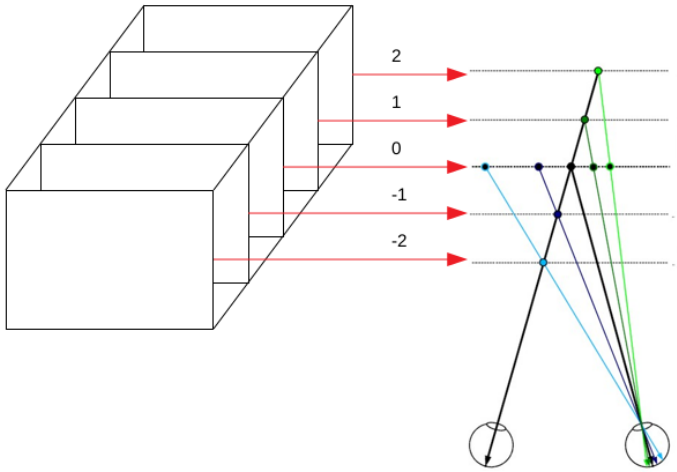


Figure 2.12: Disparity from planes of different depths: *Far binocular cells* respond to disparities on planes 1 and 2. *Near binocular cells* respond to disparities on planes -1 and -2. *Tuned zero cells* respond to disparities on plane 0 (*fixation plane*) [38].

This subdivision of the three-dimensional space has fundamental importance for the development of the computational model described in the chapter 3. Assuming that between the two stereoscopic representations of the same point, only its location changes, its position in the three dimensional space can be determined from the evaluation of the difference between the position in the left and the right view. The displacement is proportional to a specific value of *retinal disparity*. As visible in figure 2.8 and in figure 2.12 a bigger retinal disparity value , considered along the horizontal direction, indicates a point that is situated on a plane (figure 2.12) closer to the observer than the *fixation plane*. Bigger *retinal disparities* indicate smaller distances from the observer, while smaller *retinal disparities* indicate bigger distances from the observer. If the disparity between the locations of the same point in the two stereo views is determined completely by the value of a horizontal shift, this can be considered as an equivalent of the *retinal disparity*. In this situation all the points lying on the same surface in figure 2.12 are associated to a specific value of the horizontal shift existing between the position of the same point in the two stereoscopic views.

2.3 Computational Models of Visual Attention

In the last three decades, with greater focus in the last 10 years, computer vision techniques development and growth of interest in robotics, have caused that even the engineers began to deepen the study of mechanisms able to select the most relevant

information within the large amount of visual data. A more technical-engineering approach has been combined with previous physicians and biologists studies. This led to the construction of new computational models of visual attention less focused on the understanding of human perception, having the purpose of improving existing vision systems. Most of the algorithms inspired buy human visual system designed with an engineering objective, have a similar structure adapted from psychological theories akin Feature Integration Theory (see section 2.3.2) or Object Based Attention (see section 2.3.3), but present different ways to implement the details [11] [10]. The general structure of a computational visual attention model includes some basic steps that are repeated more or less in all the algorithms:

1. image or image sequences given as input;
2. several features are computed in parallel;
3. features conspicuities are fused in saliency map;
4. the maxima in the saliency map are investigated by the focus of attention (FOA) in order of decreasing saliency.

While *bottom-up* saliency is a combination from different feature channels extracted from the image itself the *top-dpwn cues* may influence the processing at different level (see section 2.1.3 for further). For a general overview of this structural structure see figure 2.13. Including all the types of *top-down* information in the algorithms is not possible so only a few have been simulated in some of the existing models: the most abstract ones, such as emotions and motivations, as far as we know have not been integrated in any models yet. The input image, generally can be artificially created but a good model should be able to work with natural scenes as well. From

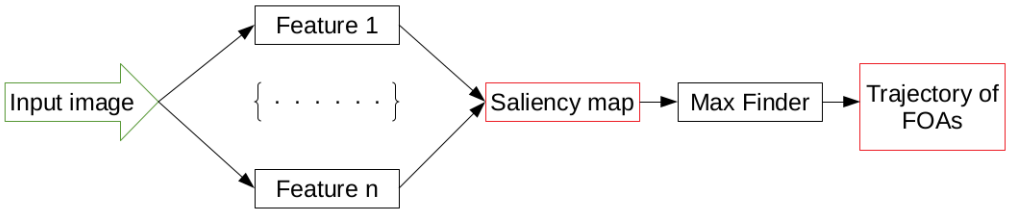


Figure 2.13: General structure of most visual attention models, the two outputs are indicated with red rectangular (figure adapted from [10] and [11]).

the focus of attention is possible to rebuilt human eye movements, both the saliency map and the focused regions can be seen as output of a computational model of visual attention.

The study described in this work considers as final output the so called *saliency map* (see section 2.3.1 for further information), all the evaluation metrics utilized

to show and evaluate algorithm performances compare this *saliency map* with a ground truth map (fixation density map estimated using an eye-tracker). For that reason this work does not present any deep explanation about the investigation step conducted by FOAs.

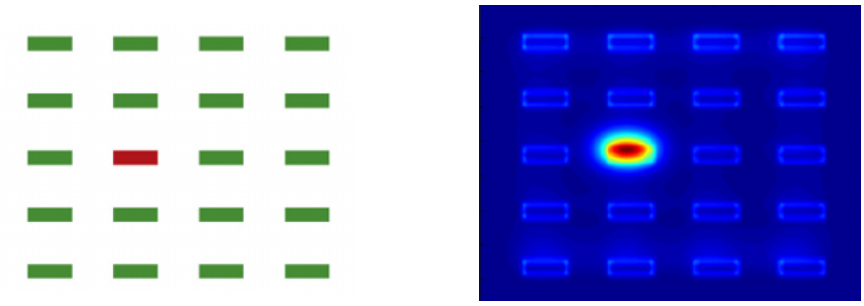
2.3.1 Saliency Map

Looking for the definition of *visual saliency*, it is found that the most used description and, in my opinion, the most explanatory for this mechanism, was provided in 2007 by Laurent Itti and is shown below in its original version.

"Visual saliency is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention."

This definition is valid to explain saliency (or salience) mechanism for each of the five senses and has been widely discussed in previous sections, now it is necessary to provide a method for estimating it at computational model level. For that reason the *saliency map* is always the first output given by all the computational model of attention developed in the last 15 years. The *saliency map* is a two-dimensional map of scalar values that integrates information from the single feature maps elaborated in parallel. It gives a description of salient region locations in the visual field and its most active regions indicate the position towards which attention will be turned [27]. The *salience map*, intended as the output of a visual salience modeling algorithm, is an image composed of pixels whose numerical value is proportional to the salience of the pixels themselves. If the image represents a realistic scene, the latter, hopefully, will attract the attention of the viewer respecting, more or less faithfully, what indicated by the map. Figure 2.14 shows an example of saliency maps evaluated from a generic algorithm, whose type and functioning do not interest us for the moment, starting from both and artificially created image and an image representing a realistic scene. Figure 2.14 has the only purpose of showing as a saliency map looks like.

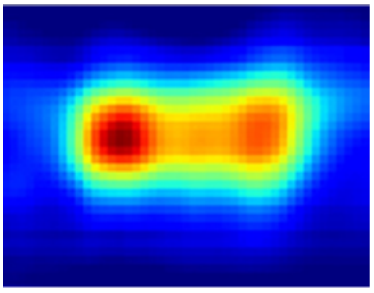
The *saliency map* is feature-agnostic [1] so a highly salient point could have been caused by a red dot among all green dots or by a vertical oriented bar among all horizontal oriented bar. Even if the *saliency map* is a good way to predict eye movement is still not clear if there is a biological correspondence to this map in visual pathways. Studying conducted on BOLD (blood oxygenation level dependent) signal [1], whose variations depend on the increase in blood flow in areas populated by nerve cells involved in the activated function, have been trying to find which area of the brain may compute a sort of *saliency map*.



(a) *Artificial image*

(b) *Saliency map*

(c) *Realistic scene image*



(d) *Saliency map*

Figure 2.14: Examples of saliency map: **(a)** Artificial Image: the salient object is expected to be the red bar (pop-out effect) **(b)** Saliency map relative to the Artificial Image: pixels numerical value (therefore their relative saliency) grows going from blue to red, red zones are the most salient ones **(c)** Realistic scene image: image representing a car on a city street **(d)** Saliency map relative to the realistic image:pixels numerical value (therefore their relative saliency) grows going from blue to red, red zones are the most salient ones.

The map can be seen ad a *retinotopic map* that is the orderly and punctual projection of the retina on the higher encefalic centers (found in mice, primates and men). Visible word is systematically mapped from the *retina* up to the higher levels of visual pathways (see section 2.1.1), regions of each area of the brain involved are projected into the following one as if the were overlapping, following proportions due to the different dimensions of the body districts involved. In the *retina* the map is precise and well-order while in the *visual cortex* it is partially distorted because here the *fovea* is represented by a bigger neurons poulation. An important demonstration of this theory based on overlapping map, is the fact that if some area of the *visual cortex* get damaged, the individual shows local blindness as if the correspondent part of the *retina* had been damaged. Different roles of different areas in visual pathways computed different kind of maps: *subcortical pathway* seems more involved in the computation of feature-specific maps, both led by *top-down* and *bottom-up* stimuli while *cortical pathway* may compute feature-independent maps, both led by *top-down* and *bottom-up* factors (figure 2.15).

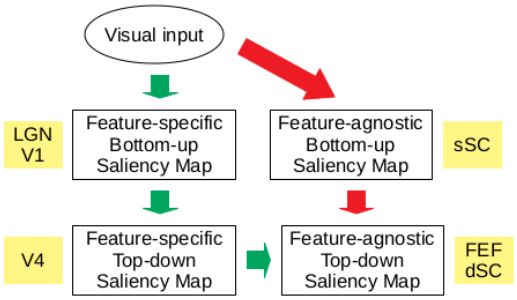


Figure 2.15: Cortical (green arrows) pathway and subcortical (red arrows) pathway for saliency computation proposed by [1] (LGN: Lateral Geniculate Nucleus, V1-V4: visual cortex, sSC: superficial Superior Colliculus, FEF: Frontal Eye Field, dSC: deeper Superior Colliculus).

2.3.2 Feature Integration Theory

In 1980 Treisman and Gelade [35] were able to explain the mechanisms of *feature search*, *conjunction search* and the differences between them, introducing a theory of *visual attention* (section 2.1.4 and figure 2.5) based on the early, automatic and in parallel perception of the features that characterize objects in the visual field. *Feature search* takes place when a target differs from the distractors, in the scene, only because of one single feature (*pop-out effect*), while in *conjunction search* the target is defined by a combination of more than one features. This theory is called *Feature Integration Theory* (FIT) and is one of the most influence theory by which many computational models of visual attention have been inspired. *Feature Integration Theory* states that the first steps of visual analysis are controlled by receptors that respond selectively to some features of the image (visual scene), each feature would be mapped in a different area of the brain. Therefore, in contrast with Gestalt psychology [21] (better explain in sections 2.3.3), here the single parts of an object (its features such as color, orientation and intensity) in the visual scene, are perceived before the whole entire object, be able to see the object required focused attention that should be a subsequent step to the first *preattentive step* [40] [27].

Models based on FIT are basically divided in two steps (figure 2.16) that takes place in the following specific order.

1. **Preattentive Stage:** features of the scene are analyzed separately while observer is still not aware of the object (parallel search).

2. **Focused Attention Stage:** features are combined to perceive entire objects (search is performed serially).

The experiments conducted by Treisman and Gelade [35] showed that if a target object is characterized by a unique feature (*feature search*) it is rapidly identified,

but if it is defined by a combination of non-unique features (*conjunction search*) then focused attention is required to bind them into a single object and the search must be performed serially. Color, orientation and intensity are good feature for *feauture search*.

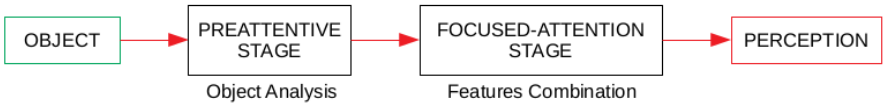


Figure 2.16: Stages of Feature Integration Theory. Reproduced from [40].

One of the most influential saliency model based on FIT is the one proposed by Itti, Koch and Niebuhr in 1998 [15], which in turn is inspired by the formalization of the FIT proposed by Koch and Ulmann [18] already in 1985 [27].

Itti, Koch and Niebur Saliency Model

Itti et al. model [15] is a derivation of Koch and Ullman model proposed in 1985 without being implemented based and inspired by *Feature Integration Theory*. To Koch and Ullman model, we own the structure that is still nowadays the skeleton of most of the visual attention models, summarized at the beginning of the current section. The model proposed by Itti et at. is one of the first that uses *Image Pyramids* in order to guarantee scale invariance to the entire algorithm. Figure 2.17 shows an overview of the model, the following list is a short description of the main steps of the model visible in figure 2.17 as well.

1. **Input image given to the algorithm:** generic image is supposed to be given as the input of the algorithm.

2. **Input image decomposed in feature channel:** following the Feature Integration Theory input image is decomposed in three feature channels: color, intensity and orientation. Individual channels are created by working on the input image (e.g. filtering) or by performing specific operations on the individual layers of the RGB image.

3. **Image pyramids:** each image of the individual channel is filtered by low-pass filter and sub-sampled in order to create 8 levels Gaussian Pyramids [44];

4. **Center-surround mechanism:** using images at different scale the model mimics the *retinal* receptive fields 2.2 (found in some LGN and V1 cells as well).

5. **Normalization:** it is important to guarantee that important conspicuities of the individual feature maps have the right influence in the calculation of the

final saliency map. All the feature maps are normalize within a range of value and are weighted proportionally to the difference between the maximum value of the map and the average value of the local maxima (previously found), In this way eventually *pop-out effects* can result in the global *Saliency Map*.

6. **Across scale addition:** in order to sum up together all the contributions, feature map pyramids (all the algorithm works with image pyramids since the first step, to guarantee scale invariance) have to be scale to a common level.

7. **Global Saliency Map:** the global *Saliency Map* is calculated as the average of the three different feature maps.

8. **Winner Take All and Inhibition of Return:** this last step to select the *focus of attention* was very innovative and inspiring for all the models to come but, as anticipated, our study is limited to evaluate the models observing the first output so the Saliency Map.

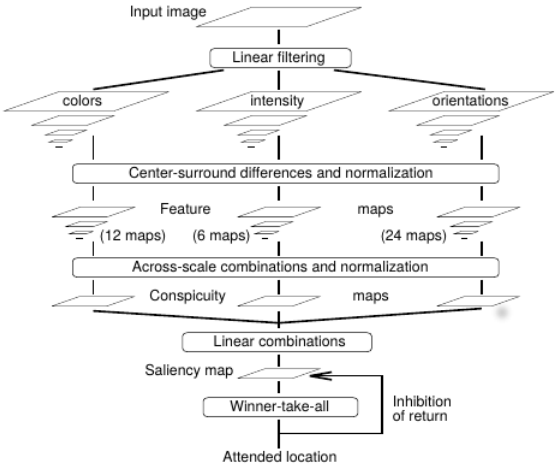


Figure 2.17: Overview of the Itti, Koch and Niebur feature based saliency algorithm. Reproduced from [15].

This model uses biologically plausible computational mechanisms and is able to perform human *feature search* (good response to the *pop-out effect* [15]) and predict human eye fixation better than chance [23]. It is still of fundamental importance and inspiration for the newest algorithms for visual salience. In the newest models of visual attention, normalization techniques as well as features number, choice and extraction have been changed but they still incorporate the concept of feature contrast and uniqueness as Knoch and Ullman proposed in 1985.

2.3.3 Object Based Visual Attention

At the same time as the birth and development of Feature Integration Theory, around the second half of the twentieth century, the first scientific publications on different object-based attention theories appear. These theories are all based on the idea that objects in the visual field are the units on which visual attention mechanism operates. According to this hypothesis, in a preattentive stage, visual scene is segmented into perceptual units (or objects) and then focal attention consists in a more detailed analysis of particular objects. The preattentive stage is a parallel process between several different objects while focal attention stage is serial because it is not possible to see too many objects at the same time. Such units (or object) can be seen as the product of perceptive grouping laws formulated within the Gestalt psychology [21] (figure 2.18).

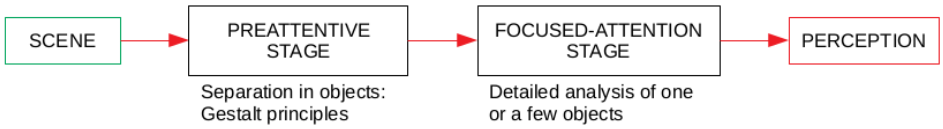


Figure 2.18: Stages of Object Based Visual Attention theory.

In contrast to feature integration theory and feature based attention, Gestalt psychology theory states that the whole of an object is perceived before its individual features [27] [21]. Since its birth it have developed surveys on learning, memory, thought and social psychology. Its founder idea is that the whole is different from the sum of the individual parts, so it is not correct dividing human experience into its elementary components. Instead, it is necessary to consider the whole as a super-ordinate phenomenon with respect to the sum of its components. Referring to visual perception in particular, the ability to perceive an object must be search in an organization headed by the nervous system instead of an image in the *retina* [41].

Gestalt Psychology

According to Gestalt psychology visual scene is divided into:

Figure that grabs attention.

Ground that is the background, where the figure sits.

An object can exist both as a *figure* and as *ground* but not at the same time, as it is clear from the example of the Rubin's vase illustration (figure 2.19): the figure can be seen as a white vase on a black background as well as two black human faces looking each other on a white background, but never both.

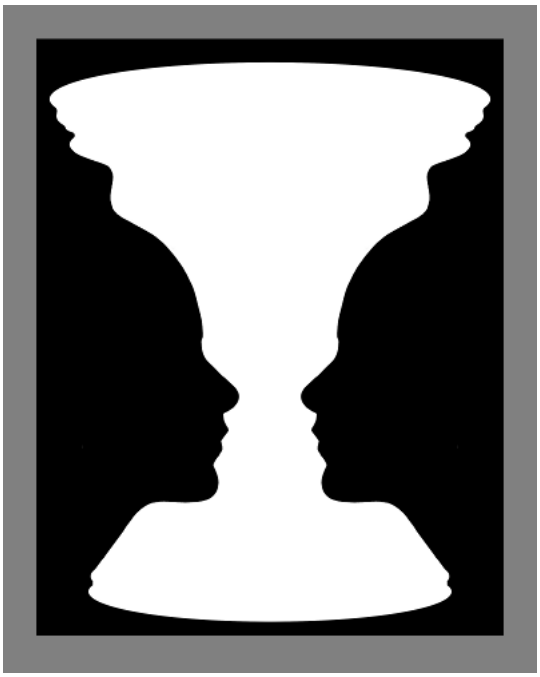
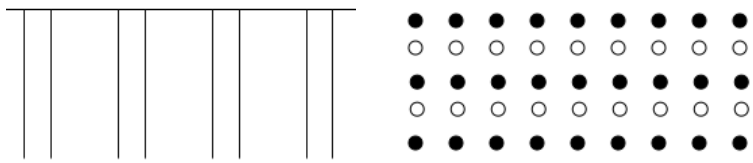


Figure 2.19: Rubin's vase illustration.

Besides figure-ground organization Gestalt psychologists have demonstrated that visual perception is based on other organizational trends that cause human perception to group together different items in the visual scene. These organizational trends are experimented in every-day life and are the fundamental hypothesis to start thinking about real word as a composition of perceptual objects. In fact human beings seem to be able to add together several elements unconsciously and divide the scene into macro organizational structures. The principles proposed by Gestalt psychology at the foundation of this hypothesis are listed below.

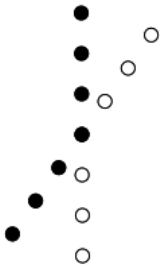
- **Proximity:** objects that are close one to the others, are seen as an unique one while elements that are distant from each other are not collected together.
- **Similarity:** tendency to gather similar elements.
- **Continuity:** elements perceived as parts of a coherent and continue ensemble.
- **Closure:** if the object is incomplete but the shape that is present is enough the whole shape can be perceived.
- **Common fate:** if the objects are moving the ones with a coherent movement are grouped together.
- **Simmetry:** objects tend to be perceived as as symmetrical shapes formed around their center.

Figure 2.20 illustrates shows some examples to better understand these concepts.

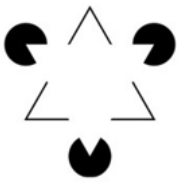


(a) *Proximity*

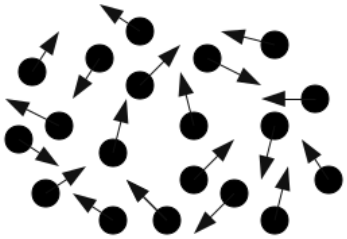
(b) *Similarity*



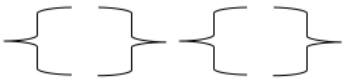
(c) *Continuity*



(d) *Closure*



(e) *Common fate*



(f) *Symmetry*

Figure 2.20: Gestalt principles of grouping: **(a)** Proximity: the image is perceived as 4 narrow columns rather than 3 wider columns **(b)** Similarity: we tent to see this figure as black lines and two white lines **(c)** Continuity: two lines crossed to form an x are perceived as opposed to two colored angles, so continuity is stronger than color similarity **(d)** in the image two triangles are seen even if one is perceived only by the presence of its corners (e) Common fate: arrows indicate the movement of each black circle in the image and the one which are moving in a common circle shape tent to be group together **(f)** Symmetry: objects are perceived as 2 symmetrical shapes.

Integrated Competition Hypothesis

The Integrated Competition Hypothesis is an object-based attention theory elaborated by Dessimone and Duncan in 1984 [9]. Starting from some experimental

results, they showed that, for a generic observer, reporting two proprieties of the same object is not more difficult than reporting only one single propriety, while reporting two features of two different objects in less effective than reporting one single object feature. They deduced that focal attention mechanism is oriented towards an object seen as a whole, between different objects there is a real competition to grab focal attention. If attention is focused on a particular feature of an object, it will enhance the processing of all the features of the object.

Sun and Fisher [32] [33] have developed a visual attention model based on this theory. They used the concept of competition between different units in the visual scene but replacing object-centered attention mechanism with a grouping-centered attention mechanism using the principles of Gestalt psychology [9]. A grouping involves one or more objects, features related to the object(s) and its or their location(s) and, as it is in the Integrated Competition Hypothesis, if a grouping grabs focal attention everything that composes it becomes important. Although based on biologically motivated theories it uses machine vision techniques: main algorithm input is given as a so called *foaveated image* obtained from the elaboration of an image collected using a retina-like sensor. Therefore the model does not provides insight into biologically mechanism which have the task of organizing the perception of a visual scene [33].

2.4 Coherence Theory and Proto-Objects

Coherence theory is an alternative hypothesis for object based attention and since it includes the definition of *proto-object*, fundamental concept for the understanding of the visual attention model described in the following chapter, it was decided to reserve for its description an entire section, separating its explanation from the others included in the generic subsection concerning visual attention object-based model. As many of the other theories of visual attention *coherence theory* include the two main stages: *preattentive stage* and *focal attention stage*. In addition to the explanation of what happens in these stages, to fully understand the core of this theory it is important to describe what happens as soon as *focal attention* is released (figure 2.21). The perception of the scene is described as a dynamic mechanism: visual field is normally composed by volatile structures that are stabilized by *focused attention* in order to make the perception of possible changes in the structures. These preattentive structures are called *proto-objects* and have been first described by Rensink in 2000 [26] as one of the fundamental concepts of *coherence theory* elaborated by Renskin himself. This study was born from the need to explain how it is possible to grasp even only a few changes occurring in the visual field, excluding that the human brain has enough memory and capacity to include somewhere, a stable and detailed representation of the stable and detail world around us. The answer to this can be found into what it means to be attend, *coherence theory*

states that attention mechanism is largely concerns with *coherence*. Scene perception occurs through the succession of three steps (figure 2.21).

1. **Preattentive Stage:** volatile, low-level proto-objects are formed in parallel across all the visual field. A proto-object is a preattentive dynamic structure that can be complex in shape but have limited spatial and temporal coherence, it is said to be volatile because it is updated if the *retina* receives a new stimulus.

2. **Focused Attention Stage:** a small number of proto-objects are stabilized as they acquire a higher level of coherence over time and space. This happens because through focused attention some proto-objects are selected and "held". Due to acquired coherence in time, if a new stimulus occurs now in the location of a stable object this is interpreted as change of the existing structure rather than the appearance of a new one.

3. **Focused Attention Release:** when attention is released from stable objects, they dissolve back into dynamic proto-object representation and there is not or little effect of having been attended.

According to this a change in the visual scene is perceived only if it occurs where and when attention is focused. Due to the small number of items that can be attended at a specific time, the probability that a change in the scene is not seen by a generic observer is very high: change blindness occurs if any changes involve only unattended items.

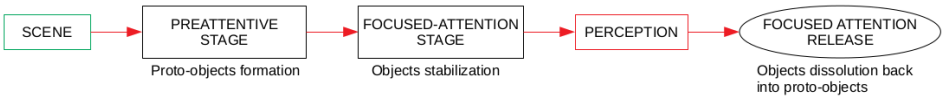


Figure 2.21: Stages of Coherence theory.

Preattentive stage or Low-level Vision

In *coherence theory* a *proto-objec* is both the highest-level output of low level vision and the lowest-level operand of attentional process that are included in high level vision. *Low-level vision* includes all the *preattentive stage* process and can be divided in three main phases that consecutively increase the level of details perceived by the observer.

1. **Transduction stage:** photoreceptor occurs, pixels level proprieties and minimal interactions are perceived. Colors and intensity can be seen and used to group areas together but no kind of complex image structure is visible.

2. **Primary processing stage:** linear filtering measures simple image proprieties as edges. At this stage nervous system can operate some easy kind of inhibition or excitation mechanism to make more complex groupings within the images.

3. **Secondary processing stage:** rapid non linear and local interpretation of the scene permits to first perceive *proto-object* structures.

All the three stages include operation that are carried out in parallel but while in the first two measurements are "quick and clean", meaning that they are simple but precise (only few errors generally occur), in the last stage interpretations are "quick and dirty" so they may not always be correct. At the end of *low-level* perception (*preattentive stage*), proto-object are accessible to be attended by *focused attention*. *Proto-objects* are volatile structures and so have limited coherence both in time and in space: they are either overwritten by new stimuli or else fade away in a few hundred milliseconds [26].

Focused Attention or High-level Vision

Focused Attention must give to the structures in the visual scene the degree of coherence needed to link them into stable larger-scale objects that are continuous in time. This is fundamental in order to make the observer able to see changes: if a new stimulus occurs now it is interpreted as a transformation of the existing structures rather than the formation of a completely new one. Therefore if a proto-object is attended it enters in a so called *coherence field*, becoming a stable objects. *Focused Attention* can be involved with the representation of only one object at the time and operates according to the following steps.

1. **Proto-object attended:** only one at the time.

2. **Link established between proto-object and the nexus:** a single structure called *nexus* is the means through which *focused attention* interacts with lower-level structure. The *nexus* contains information about the attended object (e.g. size, shape, overall color...), inside the *nexus* proprieties are computed and briefly stored.

3. **Two-way transmission of information:** information travels through the link from the proto-objects to the nexus, carrying descriptions of selected proprieties and from the nexus to the proto-objects providing them stability.

When a continuous flow is established between the *nexus* and the *proto-objects* the originated circuit is called *coherence field*

Focused Attention release and aftereffect

As already said only one item at the time can be attended by *focused attention* so if attention is shifted towards an other object the *coherence field* can not be maintained. When this happens the items that were being attended before the attention moved return to their original volatile *proto-objects* status but a short-term memory of them rests. A more in-depth explanation of what happens at the memory level is not necessary for the purpose of writing the dissertation and understanding the work done. For this reason we simply say that in reality more than short-term memory we should talk about visual short-term memory. For an object, being assisted is both necessary and sufficient to be in the visual short-term memory, so it does not seem to exist any differences between the focused attention and memory mechanisms, some studies argue instead that the memory of the previous attended objects in is lost completely. In figure 2.22 the main steps of the *coherence theory* of vision are summarized.

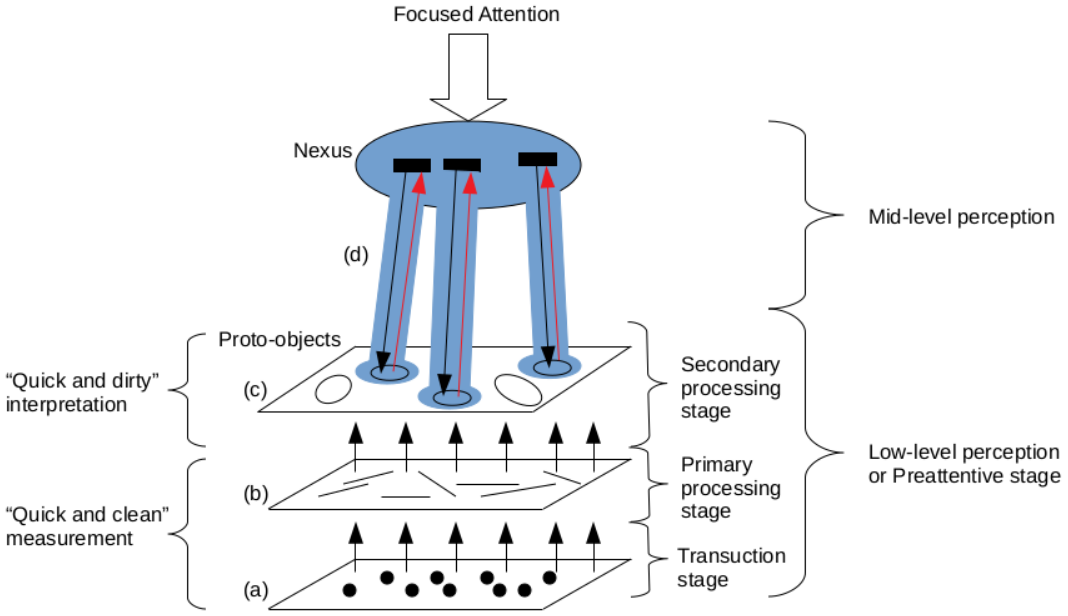


Figure 2.22: Schematic of main coherence theory process. Low-level vision composed by three main stages: **(a)** the *transduction* stage where photoreceptor operates, **(b)** the *primary* processing stage, where linear filters measure image proprieties and **(c)** the *secondary* stage of rapid and non linear interpretation that has as outputs the *proto-objects* that are the only structures from the low-level perception accessible to *focused attention*. A set of 3 *proto-objects* corresponding to objects parts create a *coherence field* (all the blue parts of the image), **(d)** are the bidirectional links between *proto-objects* and the *nexus*. Figure reported and adapted from [26].

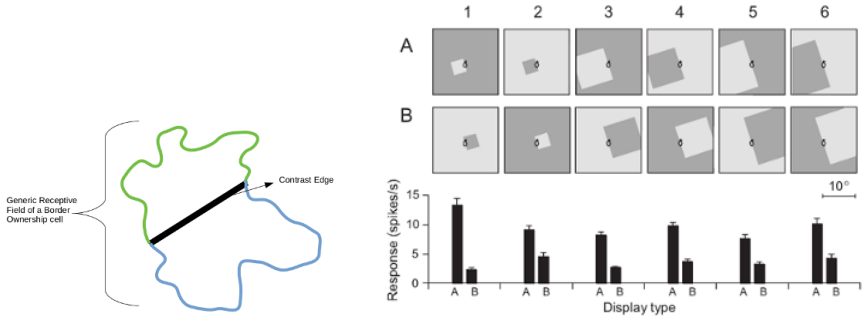
Walther and Koch Model

For the first model of visual attention that uses *proto-objects* notion, we must wait the early two thousand, when Walther and Koch [37] use it to analyze possible improvements in the mechanism, inspired by human biology, of objects recognition. The structure on the visual attention computational model is left unchanged from the one proposed by Itti et al. in 1998 [15] (see section 2.3.2). More than a *proto-objects* based model of attention, it can be seen as a model inspired by the *Feature Integration Theory* which, after having elaborated the *Saliency Map*, uses an *Inhibition Of Return* based on the proto-object notion. In fact, a part from little changes from the model proposed in 1998 (e.g. the way to build the image pyramid) the model is kept the same up to the the *Winner Take All* mechanism that permits to find the *Focus Of Attention*. In the search of the *Focus of Attention*, for each most activated point in the *Saliency Map*, the algorithm looks for the single feature map that most contributed to the activation of that specific point. In winning feature map the shape of the *proto-object* at the most activated location, is calculated spreading the activation following a contiguous 4-connected neighborhood of above threshold activity. The model just described shows that salience based on the concept of proto-objects can improve the performance of a recognition algorithm inspired by biological mechanisms. However it does not explain how *proto-objects* perception changes the model based on *feature integration theory* and, moreover, the way it uses the single feature maps to define the shape of a proto-object does not have a real response in human biology.

2.4.1 Border Ownership

Gestalt psychologists were the first to understand that to perceptual organize the scene an observer should summarily understand what is foreground and what is background solving the problem of occlusion between objects. The borders are the separation between different items in the scene so I can well seen only if these border are correctly assigned to the objects. Rubin vase in figure 2.19 is an example of a picture artificially created in a way that the whole scene changes, maintaining a meaning and global coherence, following the way borders are assigned to items [30]. The unstable perception of this particular image with its own specific time to change from one to the other interpretation, shows that in the brain there is a neural substrate involved in the interpretation of the visual scene. We generally refer to the *figure ground organization* as the capability we have to distinguish figures from the background and, in a more complex scene, this translates with the ability of discerning many different levels of occlusion in the three-dimensional scene. Perceptual organization of the scene can be explained through the concept of border-ownership with which we mean the assignment of a boundary to one of the two parts that it separates. Neurons that are activated when a generic border

is identified in the scene have been found in *extrastriate cortex*, particularly in V2, and more important, it has been discovered that their firing rates change if the side where the figure that owns that specific border changes. In figure 2.23 are reported experimental results that show how the firing rate of a specific border ownership cell in monkey decreases or increases changing the position of the item located in the visual field. Summing up, each time that in the scene there is a line dividing a generic figure in two sides, the *receptive field* of a border ownership cell is recreated (see section 2.1.1 for the definition of receptive field), than a border ownership cell responds differently depending on which side of the receptive field contains the figure (figure 2.23). Border ownership signal coding in human brain



(a) *RF of a BO cell* (b) *Response of a BO cell*

Figure 2.23: Border Ownership (BO) cell response. **(a)** Generic receptive field (RF) of a BO cell: an edge divides a generic figure in two parts (the green one and the blue one) and the one that owns the border can be in both sides, depending on its position the firing rate of the Border Ownership neuron changes. **(b)** Response of a BO cell in monkey V2. Rows A and B show the stimuli and the RF of the B0 cell (black ellipse) . Bar graph shows mean neuron firing rate, neuron has a preference for the left side (higher firing rate when the square is on the left side of the RF)

has been studied during the last 10 years without reaching a unique common model. All the models are supported by neural theory of information transport at neural level and apparently they seem to be equally valid [30], but for our purpose to build a proto-object based saliency model the most suitable seems to be the *feedback model*. Moreover, the model described in this work is an extension of the model proposed by Russel et al. in 2014 [28] so, as the previous model, it is based on an algorithm that reproduces the *feedback model* as faithfully as possible to human biology. Visual research experiments on macaques have shown that border-ownership signals can appear even after only $20ms$ from the occurrence of the stimulus and this speed is too fast for being explained with the lateral propagation through *primary visual cortex* [27]. Despite many of the studies mentioned above have recorded the neuronal activity in the macaques V2 cortex, it is important to remark that the presence of *Border Ownerhip* cells was also found in humans both through psychophysical

studies and border-ownership-selective BOLD signal recording [30].

2.4.2 Grouping Mechanism and Craft et. al Model

Craft et al. in 2007 [7] were the first that purposed an alternative to the lateral propagation for border ownership signals supposing that border ownership cells have a close and intense exchange of information with *grouping cells* populations. The *grouping cells* are situated at higher levels of visual cortex and communicate with border ownership neuron through withe matter projections that can permit to have such a high speed communication. Every time an object is present on the scene for each contour line two types of border ownership cells are activated and compete for the ownership of the given border, this cells are connected with appropriate *Grouping Cells* which therefore are activated and play a fundamental role in the competition between the *Border Ownerhip* cells. *Grouping Cells* integrate information about the contour and through a feedback mechanism enhance activity of the *Border Ownership* cells which code for the figure: feedforward synapses from *Border Ownership* cells excite *Grouping Cells* while feedback connections facilitate *Border Ownership* cells activity (see figure 2.24 for the *grouping cells* mechanism). Each *grouping cell* targets many neurons at the lower level (all the *Border Ownership* cells that are consistent with the same object in the scene) so those neurons show an increase in synchrony when the common *Grouping Cell* is activated. In reality we always refer to a population of *Grouping cells* better than a single *Grouping cell*. The model of visual attention proposed by Craft et al. [7] proposed the concept of *grouping cells* as the way the brain has to integrate object features into tentative proto-objects without needing to recognize the object [28] and using the principle from Gestalt theory, in particular continuity, closure and proximity (figure 2.20). The model aims to perform specific cases of *figure-ground segregation*. Figure 2.25 shows the how grouping cells model integrates perfectly within the Gestalt theory and how this latter, vice versa, finds a biological confirmation in the grouping cells model.

The algorithm describes in the next chapter, which is the central question of this dissertation, is an extension of the attention model proposed by Russell et al. [28] which in turn is strongly inspired by the 2007 Craft model [7].

2.5 Visual Attention Models in three-dimensional space: an overview

All the models mentioned from the beginning look at the surrounding world through images from which features are extracted to analyze the visual scene trying to simulate human visual perception mechanism in the most plausible biological way. The

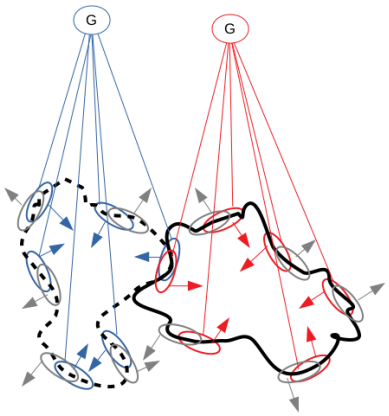


Figure 2.24: Schematic of Grouping cell mechanism. All the ellipses show the Receptive Fields of Border Ownership cells and the arrows indicate the relative preferred side, with respect to the present border. Presence of the dotted line object increase the firing rate of all the blue cells over the gray and the red ones because their receptive fields are consistent with the dashed line object so they receive feedback facilitation from the relative grouping cell while whit the solid black line object gray and blue cells receive no (or less feedback) because they are not consistent with the present object, this time red cells receive feedback.

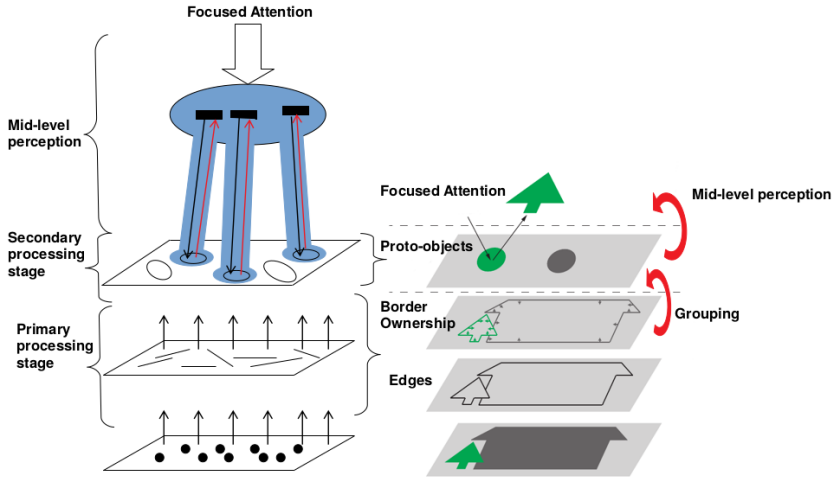


Figure 2.25: Coherence theory and grouping cells model. Each step that, according to coherence theory, leads from the presentation of the visual stimulus (visual scene) to focus attention, finds confirmation in the steps of the grouping mechanism and vice versa. Adapted from [27].

fact that none of the models taken as examples takes into account any binocular cues or, more simply, a generic depth information reflects the disproportion that exists, considering the current state of the art, between the number of computational models of "two-dimensional" visual salience and the number of models that consider

how the human visual attention is distributed in three-dimensional space. More than this, due to the relatively recent development of the proto-object concept, very few models fuse grouping mechanism and depth perception. In any case until now grouping mechanism concepts have been used to carry out a sort of figure ground organization in the three-dimensional space. Many three-dimensional models have tried to introduce information to a given two-dimensional model without modifying its supporting structure as it has already been observed that in reality depth information, although influential, drive less visual attention than features such as color or brightness. For this reason, generally, instead of using new resources, we try to adapt the already existing models, that is exactly what our model does. A problem with the three dimensional models is the difficulty to create a dataset that includes binocular cues. As explained in section 2.2 binocular cues are fundamental for human depth perception so generally a dataset for the evaluation of three-dimensional saliency model has to deal with stereo vision and disparities between the two views, that makes it difficult to create. Following the approach used to include depth perception, three-dimensional visual saliency models can be divided in three main categories [13]:

- **Depth-weighting models**;
- **Depth-saliency models**;
- **Stereo-vision models**.

The first two types of models require as input the image that represents the real scene and the corresponding depth map, while the last one gets information about depth directly from the two stereoscopic view of the same scene.

Depth Map

The depth map associated to a specific image of a scene is also an image where the pixel intensity is related to the proximity from the observer of the corresponding point in the real scene. They can be computed directly by the instrumentation (e.g. Kinect) or through stereoscopic analysis. If the first method is strictly linked to the availability of the necessary instrumentation, the second requires the development of algorithms that are not always easy, especially if a certain level of calculation precision is required.

Depth-weighting models

This category of three-dimensional saliency model requires both a two-dimensional image and depth map as inputs. The output saliency map is calculated following a generic two-dimensional model and once it is elaborated is weighted through a multiplication, using the depth map values. These models follow the basic assumption

that objects closer to the observer are more ecologically relevant than those which are more distant. In this way the computational cost does not increase and the existing two dimensional models are easy to adapt but you may not detect areas whose salience may depend only on features related to the depth of the scene.

Depth-saliency models

The Depth-saliency models as the Depth-weighting ones require depth map as input but differently from previous models this time the map is used to extract additional information in order to create a saliency map from the only depth features. This mean that the model has the two following intermediate outputs.

- **A two-dimensional saliency map:** elaborated from the ordinary features (e.g. color opponency, intensity, orientation of the objects...) extracted from a generic RGB image. It does not take in consideration the scene development in the third dimension.
- **A depth saliency map:** includes only information from depth features extracted from the depth map given as inputs.

These two intermediate outputs are then combined generally through an addition weighed by multiplicative factors. All these operations introduce a real change to the original model increasing the total computational cost and the model requires the additional input of the depth map.

Stereo-vision Models

This models replicate the most efficient solution of human being and, in particular, his visual system, to extract information about the distance of the objects in a scene, making a kind of comparison between the projection of the right retina and that of the left retina, of the same scene. Depth maps are not required as inputs but both the stereoscopic views are needed. Generally these models have a higher computational cost than the other two but somehow are considered more biologically plausible.

Among all the models we must focus on a model proposed by Hu et al. [13] whose structure is very similar to the one proposed in this dissertation. They added a parallel channel to the same proto-object based visual attention model for which we implemented the extension to include depth information. This channel takes as inputs a depth map and elaborates its own saliency map. The global structure is left unchanged as the two-dimensional algorithm that, meanwhile, elaborates its saliency map, once the two saliency maps are obtained they are combined together. This model shows little but statistically significant improvements introduced to the "only" two-dimensional proto-object based saliency model, by adding figure ground organization in three-dimensional space.

Starting from these improvement we decided to work in the same direction but including the concepts of stereo-vision in order to free the model from the need of a pre-computed depth map.

2.6 Aim of the Study

In 2014, the publication of a work [28] done at the Department of Electrical and Computer Engineering of the Johns Hopkins School University, about the development of a biologically plausible model of proto-object based visual saliency, establishes an important turning point for open questions on visual attention theories based on the perceptive organization of the scene in structures that can drive focused attention. Using basic computation mechanisms with known biological correlates, the model and its promising performances make us firmly believe that the growing belief that attention is a based object is more than founded.

Therefore our study is born with the purpose to extend this proto-object based saliency model in order to increase its performance. In particular we aim to add to the model the contribution of depth information, following the rules of human perception. Our work is not the first to attempt to extend the model in this direction but, for the fist time, instead of using a pre-computed depth map as in previous works [13], the model uses stereopsis directly following notions we have from human neural system.

Two views of the same scene, from left and right retina are given as input to the algorithm. They are identical except that the retinal location of the same image point is shifted between the left and the right view, the value of this shift is inversely proportional to the distance of that point from the observer.

We build the current model starting from two fundamental assumptions.

1. **Grouping cells mechanism is activated to enhance the activity of Binocular Neurons:** the fact that Grouping mechanism interests the organization of the visual scene in three-dimensional space has been observed in previous studies [25] [13], the way with which this happens is still matter of study. We consider that the activity of border ownership cells and related grouping cells population, takes place within a specific depth plane (considering frontoparallel planes parallel to the head coronal one). This means that grouping mechanism starts only once binocular neurons have dived the scene in parallel planes.

2. **Pop-out mechanism is also related to the distribution of objects in the three-dimensional visual scene:** we built the current model starting from the assumption that, as it is with other features such as orientation and intensity, discontinuity of depth attracts human attention more than regions of a scene that have constant depth.

As far as we know in literature there is no model that has already tried to study the influence of depth discontinuity in the visual scene, trying to predict how these can change the focus of human visual attention. On the contrary some, albeit not many, models have tried to use the concept of proto-object and the grouping algorithm to distinguish the foreground from the background and recognize an extended object in three-dimensional space. What this study proposes, on the other hand, is closely linked to depth discontinuities search. Assuming that from one stereo view to the other, pixels representing the same object have all undergone the same horizontal displacement or otherwise include within a narrow range, the algorithm works by looking for objects interested by "uncommon" displacements (disparities). These objects are hypothetically lying on a plane of depth in which few or no other objects are present. This condition is assumed to give rise to the pop-out effect making those objects more salient than others. The extension of the proposed original algorithm is developed through the addition of a new feature channel that assumes as input a couple of stereoscopic images, with the purpose to see how the evaluated final saliency map changes if we include the search for perceptual objects in three-dimensional space. Once the algorithm has been written, first of all it is necessary to make sure that the introduced feature channel is able to do what it was thought of: finding depth discontinuities. Only after having verified this we can proceed to see what happens by adding this new information to the search most salient areas in natural scenes.

Chapter 3

Methods

The previous chapters were extremely important to introduce what is the core of our study. Pretending to build or only to understand a computational model for human visual saliency without having even only few and basic notions about human visual system, both about anatomy and physiology, is not possible especially if we are interested in elaborating algorithms that are as much as possible biologically plausible. Remembering that the factors which drive visual attention can mainly be divided in two categories, bottom-up factors specific for the present scene, independent from the internal organism state, and top-down factors that involve the internal state of the organism at that specific time, the top-down factors are not considered in the present algorithm due to the difficulty to predict, measure and quantify them.

In order to contextualize the present work it is also important to know which are the major theories from which the computational models of visual attention draw inspiration. In particular, the model we approach to describe is an object-based one (see chapter 2 for more detail information) so it refers to both psychophysical and neurophysiological studies which show that attention depend on the scene organization into perceptual objects as well as image features. According to Rensink's coherence theory, the perception of these perceptual objects is generated when an observer is looking at a proto-object (see section 2.4), a dynamic structure which exists as a stable object only when it is attended while normally, when is not attended, exists in an indefinite state where it returns as soon as attention is released.

As already mentioned in the first two chapters, the state of the art of visual attention computational models, despite their study began only at the end of the last century, is really rich in algorithm examples. Many of these follow the structure proposed by Itti et al. in 1998 [15] (shown in figure 2.17) but they all differ one from each other for details or for the addition of information to be taken into account. The model described here, as well, draw inspiration from the Itti et al. main structure but adding the biological concepts of Border Ownership Cells and Grouping Cells as in Craft er al. model [7]. The basic structure, described in the next three sections,

is left unchanged from the one described in the study proposed by Russel et al. [28], the whole model was conceived and realized entirely in the department of Electrical and Computer Engineering at Johns Hopkins University in Baltimore, Maryland, United States of America. This model includes a new *Grouping Mechanism*, inspired by Redskin's therory [26] and Craft et al. Model [7] in a main framework similar to the one proposed by Itti et al. in 1998, showed in figure 2.17 and that, as already said, is still the most influential of all the saliency models. The information from the input images is separated in feature channels and *Grouping mechanism* is applied separately to each one of them. In fact studies have demonstrated that Border Ownership Cells are able to recognized and assign an edge to the correct figure even if contours are defined exclusively by a characteristic of the image: an object can be perceived only from color or brightness variation with respect to its surroundings [13] [25].

Recent researches have demonstrated that Border Ownership cells can work with stereoscopic edges when monocular cues are missing and parallel studies based on experimental results, as already mentioned in section 2.2.2, showed that human visual research, and so that human visual saliency, is strongly influenced by the distribution of the objects among depth planes, that are parallel to the head coronal one (see previous chapter for further information). Based on this and on the knowledge of the existence of specific neurons for the coding of binocular disparities (section 2.2.1), next chapter describes an extension of the proto-object based saliency model to the third dimension, using models neurally inspired disparity units. This extension is conducted by introducing a fourth channel: the *Disparity Channel*. The *disparity channel* is build from the assumption that, as it is with other image features such as orientation, intensity etc, discontinuity of depth attracts human attention more than regions of a scene that have a constant depth, and that disparities between two stereoscopic views can be used to estimate such discontinuity. The model includes three main steps, two conducted in parallel and the final one done after the others.

1. **Evaluation of a proto-object saliency map in two-dimensional space**: two dimensional saliency map gives as output.

2. **Evaluation of the *Disparity Channel* contribution**: disparity saliency map gives as output.

3. **Linear combination of the two saliency maps**: three-dimensional saliency map gives as output.

Due to the fact that the algorithm works with stereopsis principles a stereo images couple is required as input, one of the two views is used to evaluate the two-dimensional saliency map following the Russeal et al. mechanism while both of them are required to evaluate the disparity saliency.

Our work focus on elaborating a correct way to give disparity information to the grouping algorithm so that it can be apply following human biology. Figure 3.1 briefly summarizes the steps of the whole algorithm, better explanation on the new *disparity channel* are given in the next sections and the "image pyramid" step indicates that, in order to provide scale invariance to the proto-objects search, the image is down-sampling several times consecutively so as to create the pyramid. The model was implemented using MATLAB2017b (Mathwork, Natick, MA,USA).

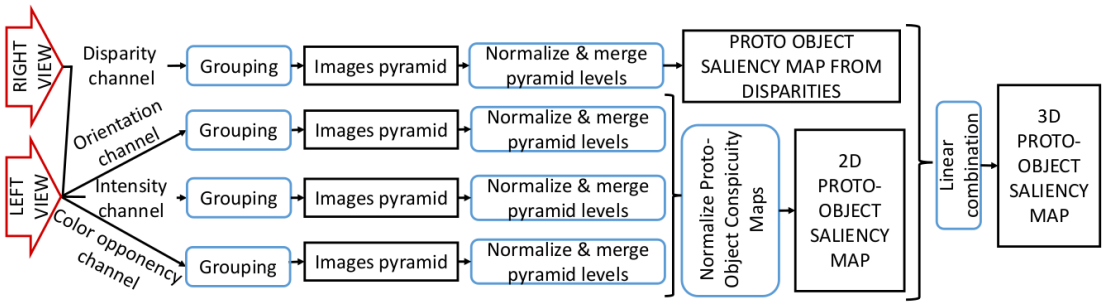


Figure 3.1: Overview of the stereo-vision model of proto-object based saliency. Color, intensity and orientation channels are identical to the model from Russel et al., see figure 5 of [28]. These three channels are used to evaluate the two-dimensional saliency map.

3.1 Two-dimensional Proto-Object based Saliency: Russel et al. model

As already mentioned this work would have not existed without previous study conducted in 2012 by Alexander Russel and his team [28] [27] on proto-object based saliency. Whit that in mind it is really important to have an overview of its base structure, it should also be considered the fact that much of this structure repeats itself identical in the elaboration of the information from the *Disparity Channel* that is the real core of this work.

The algorithm is composed by two main mechanisms.

- **Grouping mechanism:** permits to find proto-objects location and spatial scale within the input image, saliency information are provided through the organization of the scene into figure and ground (find possible preattentive structures).

- **Main framework:** global structure of the algorithm (similar to the one proposed bt Itti et al. [15] see figure 2.17 and section 2.3.2 for a general view) where the grouping algorithm is placed, it permits to distinguish objects with unique characteristic in the scene from the ones that are common.

Model fundamental steps are shown in figure 3.2 and briefly summarized below.

1. **Input Image given to the algorithm:** RGB image is given as input, it can both be an artificially one, created with the only purpose to test the model or an image representing a more natural scene.

2. **Information divided in feature channels:** different information, regarding different features of the image/scene (intensity, color opponency and orientation), are extracted from the input image.

3. **Creation of image pyramids:** this step is important to guarantee scale-invariance to all the following steps. Pyramid is created for all the feature channels and it let the following search for proto-objects to be done in different scales.

4. **Grouping mecanism:** this step is well explain in the next section. It is repeated for each level of the pyramid of each feature channels.

5. **Normalization and Pyramid levels merging:** this two steps are collected together because, depending on the different channels, some normalization may be needed before the merging of the pyramid levels to a common scale. At the end of this step for each channel have been evaluated a saliency map related to the single feature.

6. **Normalization of Proto-Objects conspicuty maps:** the single feature saliency maps are summed together.

The information from the input image is first of all divided in feature channels, in the final version of the model that came out in 2014 [28] the input image is decomposed in three channels:

Intensity Channel takes into account the sensitivity of the rods (photoreceptors of the retina) to the brightness;

Color Opponency Channel take into account the sensitivity of the cones (photoreceptors of the retina) to the different wavelengths in the visual spectrum;

Orientation Channel takes into account the sensitivity of he primary visual cortex cells to the different orientations in the scene.

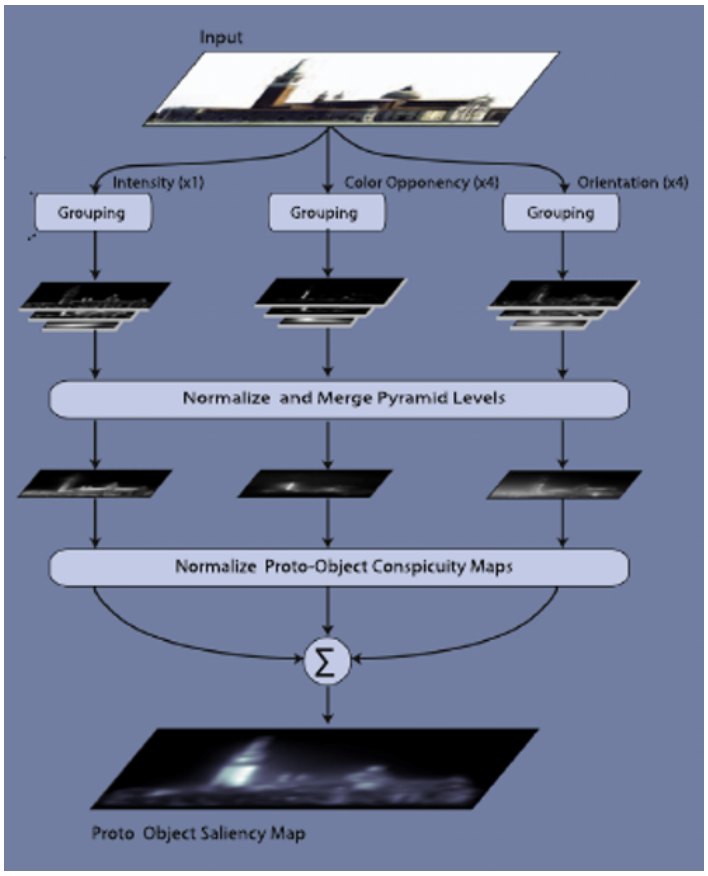


Figure 3.2: Main framework of the proto-object based saliency model proposed by Russel et al. Image reported from [28].

Image pyramid

After having extracted single feature inputs (see section 3.1.2) images are down sampled subsequently in steps of $\sqrt{2}$ using a *Bicubic Interpolation*. The number of pyramid levels can be chosen separately each time the algorithm runs, in the previous model we are discussing about, the pyramid spans 7 levels, for our simulations the number is settled to 10.

Each k^{th} level of the pyramid required two consecutive steps to be created.

1. **Image Filtering:** image from the $(k-1)^{th}$ level is filtered using a convolution kernel composed by piecewise cubic polynomials (cubic kernel, figure 3.3 **a**). So the output pixel value is a weighted average of pixels in the nearest 4-by-4 neighborhood. The convolution kernel $k(x)$ in defined as follow:

$$k(x) = \begin{cases} (a+2)\,|x|^3 - (a+3)\,|x|^2 + 1; & \text{if } x \leq 1 \\ a\,|x|^3 - 5\,|x|^2 + 8\,|x| - 4a; & \text{if } 1 < x < 2 \\ 0; & \text{if Otherwise} \end{cases}$$

2. **Real Subsampling:** the number of the pixels that compose the filtered image from $(k-1)^{th}$ is reduced by a factor of $\frac{1}{\sqrt{2}}$.

Figure 3.3 shows the Image Pyramid creation steps.

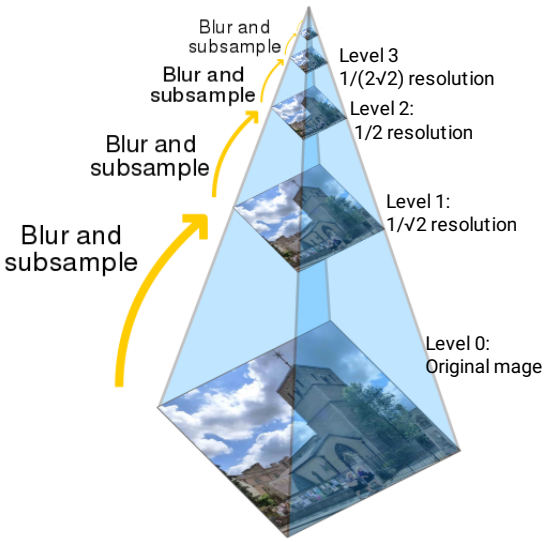


Figure 3.3: Representation of an image pyramid with 5 level. In the case of our model the blur is done through a bicubic kernel and the subsampling is done by steps of $\sqrt{2}$ as shown in the image, adapted from [44].

From this point until the end of the chapter the superscript k indicates that we are working with the k^{th} level of the image pyramid. Once the pyramid is created for each feature channel the *Grouping Algorithm* is run on each one of them. Next section is entirely dedicated to the description of the groping mechanism as proposed in the proto-object model under analysis [28].

3.1.1 Feed forward Model of Grouping

As already mentioned this part of the algorithm integrates the notion of *Border Ownership Cells* (BO cells) and *Grouping Cells* (G cells) following the model proposed by Craft et al. [7]. Each one of the following steps is repeated for each level of the *image pyramid* in each one of the feature channels, the mechanism does not have recurrent connects and all the steps that include a correlation represent neurons effect for a specific receptive field. This receptive field is described by the convolution

kernel used for the correlation under analysis. The main parts of the algorithm are shown in figure 3.4.

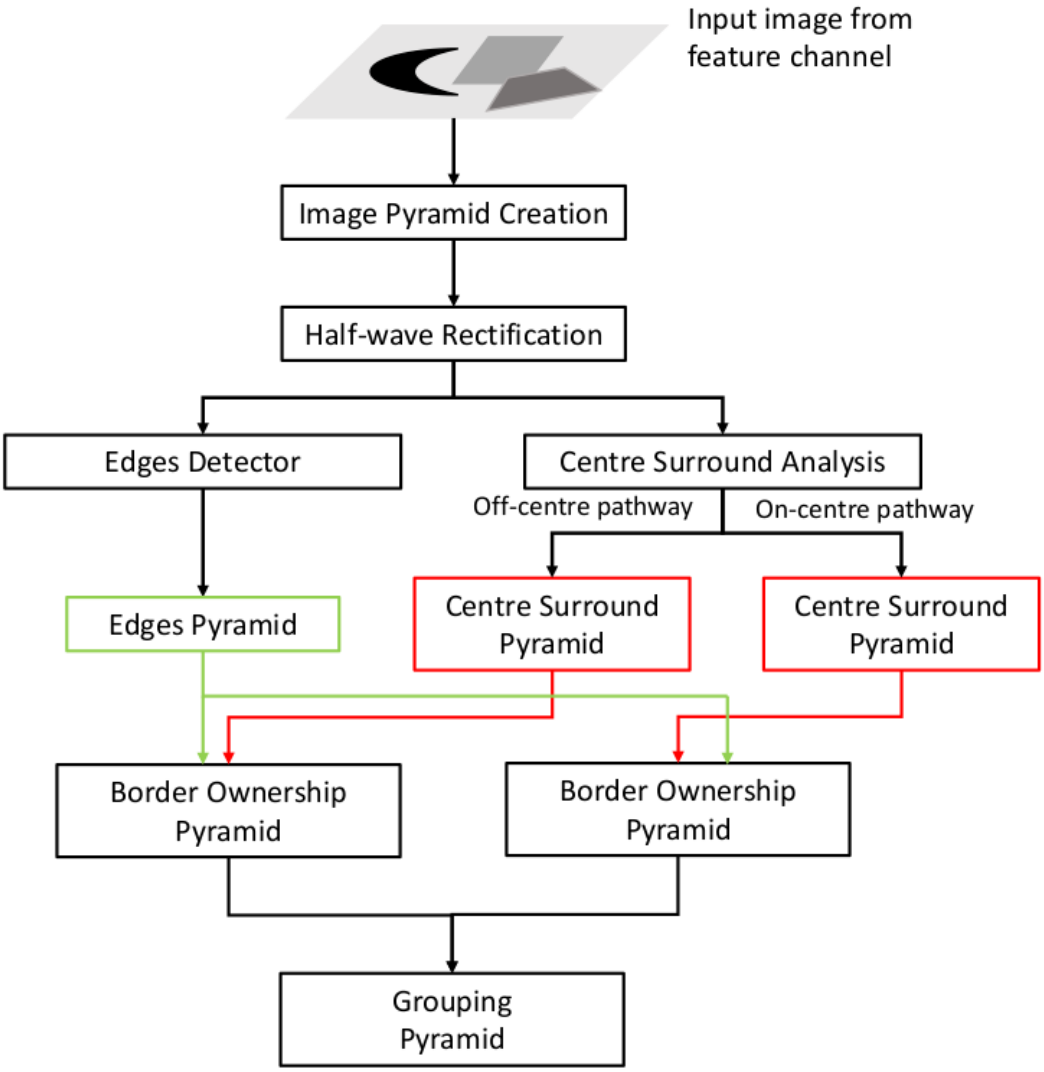


Figure 3.4: Overview of the feed forward grouping mechanism on a generic feature channel, created from figure 5 in [28].

Grouping algorithm fundamental steps are listed below.

1. **Half-wave rectification:** grouping mechanism receives input images in the range of [−1,1], negatives pixel values are present when, within a specific feature channel an opponency operation is done (e.g. color opponency channel).

Therefore considering $\beta(x, y)$ the image pyramid and $\beta^k(x, y)$ the k^{th} level of the pyramid the operations done are the following:

$$k(x) = \begin{cases} \beta^k(x, y) = 0; & \text{if } \beta^k(x, y) < 0 \\ \beta^k(x, y) = \beta^k(x, y); & \text{if } \beta^k(x, y) > 0 \end{cases}$$

2. **Edges detector:** object edges are extracted using two-dimensional Gabor filters that simulate the responses of simple and complex cells of primary visual cortex.

   Simple and complex cells are supposed to respond to a range of edges orientation. Calling θ the angle indicating a specific edges orientation, in this Russel et al. [28] use the following range of angles: $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$.

   - **Simple Cells Responses ($S^k_{e,\theta}$, $S^k_{o,\theta}$):**
     Single cell are suppose to have a receptive field similar to the kernel used for a *Gabor Filtering* that is composed by a *harmonic function* with a *Gaussian envelope*. Two types of *simple cells* are distinguished, depending on whether their receptive field is represented by a harmonic with an even or an odd trend.

     $$\mathrm{S}^k_{e,\theta}(x, y) = \beta^k(x, y) * \mathrm{g}_{e,\theta}(x, y)$$

     $$\mathrm{S}^k_{o,\theta}(x, y) = \beta^k(x, y) * \mathrm{g}_{o,\theta}(x, y)$$

     Where $*$ denotes a correlation operation, $\beta^k(x, y)$ is the image at the k^{th} level of the pyramid and the two kernels are the following.

     $$\mathrm{g}_{e,\theta}(x, y) = \mathrm{e}^{\frac{x'^2+\gamma^2+y'^2}{2\sigma^2}} \cos(\omega x')$$

     $$\mathrm{g}_{o,\theta}(x, y) = \mathrm{e}^{\frac{x'^2+\gamma^2+y'^2}{2\sigma^2}} \sin(\omega x')$$

     Even and odd Gabor filters with different orientations are shown in picture 3.5, relative parameters and the expression in table 3.1.

γ	σ	ω
0.8	2.24	1.57

Table 3.1: Gabor filter parameters.

x' and y' are the coordinates in the rotated reference frame. Their expressions are the following.
$x' = x\cos(\theta) + y\sin(\theta)$ and $y' = -x\cos(\theta) + y\cos(\theta)$.

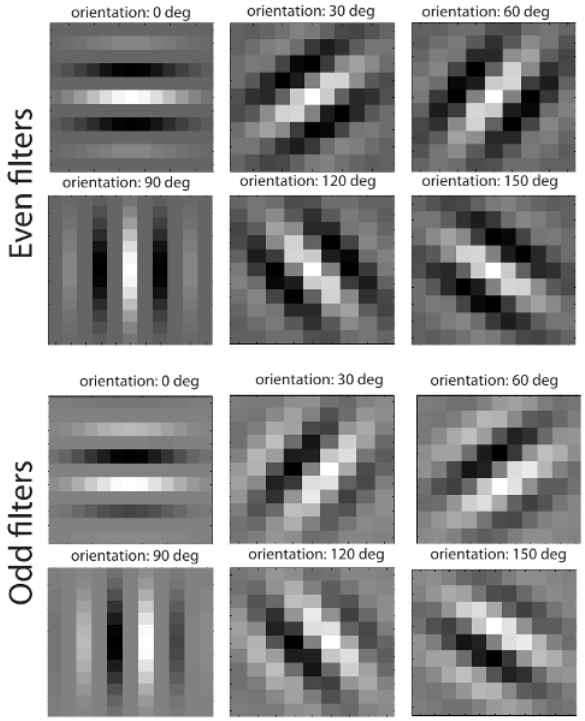


Figure 3.5: Even and odd Gabor filters. Reported from [27].

- **Complex Cells Responses (C_θ^k):**
  Complex cell's response at angle θ is calculated from a simple cell response pair as follow.

$$\mathrm{C}_\theta^k(x,y) = \sqrt{S_{e,\theta}^k(x,y)^2 + S_{o,\theta}^k(x,y)^2}$$

3. **Center Surround Analysis:** center surround mechanism is used to understand if and edge in the scene correspond to an object. Concerning a relation to the human biology, this type of information has been found in the *retina*, *lateral geniculate nucleus* and *primary visual cortex*. Following knowledge from biology and in particular notions on retinal *Ganglion Cells* receptive field (figure 2.2) the model proposes to filter images, from each level of the pyramid and for all the feature channels, using a pair of kernels that reproduces respectively the ON-center and the OFF-center receptive field. They both have circular shape divided in two antagonist concentric regions: the central one and the peripheral one (figure 2.2). The ON-center receptive field can identify a light object on a dark background while the OFF-center receptive can identify a dark object on a light background. The activity of the neurons in visual pathways that have these specific types of receptive field is computationally

simulated by the following two correlations.

$$\mathrm{CS}_D^k(x,y) = \beta^k(x,y) * \mathrm{cs}_{off}(x,y)$$

$$\mathrm{CS}_L^k(x,y) = \beta^k(x,y) * \mathrm{cs}_{on}(x,y)$$

where $CS_D^k(x,y)$ and $CS_L^k(x,y)$ are respectively the OFF-center and the ON-center pyramid, the two kernels are built using a difference of Gaussians as follows (examples in figure 3.6).

$$\mathrm{cs}_{on}(x,y) = \frac{1}{2\pi\sigma_i^2}\mathrm{e}^{-\frac{x^2+y^2}{2\sigma_i^2}} - \frac{1}{2\pi\sigma_o^2}\mathrm{e}^{-\frac{x^2+y^2}{2\sigma_o^2}}$$

$$\mathrm{cs}_{off}(x,y) = -\frac{1}{2\pi\sigma_i^2}\mathrm{e}^{-\frac{x^2+y^2}{2\sigma_i^2}} + \frac{1}{2\pi\sigma_o^2}\mathrm{e}^{-\frac{x^2+y^2}{2\sigma_o^2}}$$

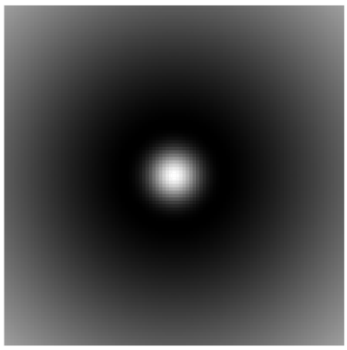
In the previous two expressions σ_i and σ_o are the standard deviations of the inner and the outer Gaussian. Before proceeding, a normalization is carried out on the pyramids CS_L^k and CS_D^k in order to promote the activity of maps with few items and to suppress the activity of maps with distractors. The normalization process is repeated identical on the final single feature saliency map and it is explained in section 3.2.6.

4. **Antagonist pair of Border Ownership:** *Complex Cells* activity and *Center Sourround* mechanism modulate each other in order to recreate *Border Ownership* cells activity. As we know from section 2.4.2 *Border Ownership* cells work in antagonistic pairs: for the same edge there are two populations of *Border Ownership* that are activated by the presence of the object in one of the two sides of the contrast edge. For example for an horizontal edge one of the two *Border Ownership* cells from and antagonistic pair is activated if the object is located above the border while the other is activated if the object is located under the line. Once the border is identified by the *complex cells*, *center surround* mechanism add information about the object itself, looking for a light object on a dark background ($B_{\theta,L}^k$) or the opposite situation ($B_{\theta,O}^k$). *Border Ownership* activity for a light object on a dark background is evaluated as it follows.

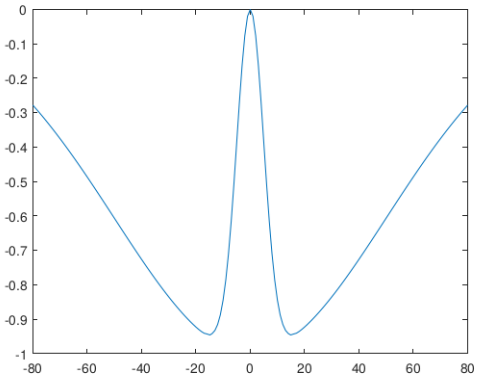
$$\mathrm{B}_{\theta,L}^k = \mathrm{C}_\theta^k \mathrm{x}\left(1 + \sum_{j\leq k}\frac{1}{2^j}\nu_{\theta+\pi} * \mathrm{CS}_L^j - \mathrm{w}_{opp}\sum_{j\leq k}\frac{1}{2^j}\nu_\theta * \mathrm{CS}_D^j\right)$$

At the same way *Border Ownership* activity for a dark object on a light background is defined as it follows.

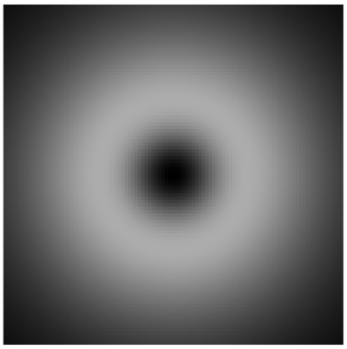
$$\mathrm{B}_{\theta,D}^k = \mathrm{C}_\theta^k \mathrm{x}\left(1 + \sum_{j\leq k}\frac{1}{2^j}\nu_{\theta+\pi} * \mathrm{CS}_D^j - \mathrm{w}_{opp}\sum_{j\leq k}\frac{1}{2^j}\nu_\theta * \mathrm{CS}_L^j\right)$$



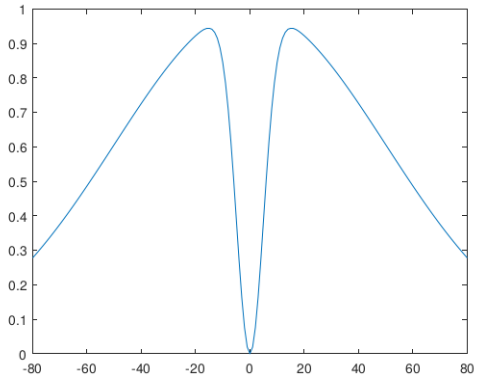
(a) *On-center operator: 2D representation*



(b) *On-center operator: 1D representation*



(c) *Off-center operator: 2D representation*



(d) *Off-center operator: 1D representation*

Figure 3.6: Center Surround operator. All the four figures have been created using MATLAB2017b and have only an illustrative purpose.

Where:

- ν_θ is the annular kernel that links *Center Surround* activity with the information of the object edges. It is generated using Von Mises distribution and normalized in the range $[0-1]$; In order to produce the desire annular receptive field eight individual kernels are evaluated for eight different angles θ and then combined. To well position kernels in the image their expression depends on the *zero crossing*; radius value of the *center surround* receptive field (figure 3.7).
- w_{opp} is the synaptic weight of the inhibitory signal from the Center Surround pyramid with the opposite polarity.
- 2^{-j} is a factor used to normalize ν_θ to have constant influence across spatial scale.

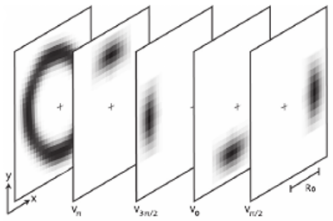


Figure 3.7: Annular kernel used to map the Center Surround cells activity to the Border Ownership cells, each kernel is generated from Von Mises distribution at a specific angle. In figure are reported kernels at $0, \pi/2, \pi, 3\pi/2$ and is indicated the zero crossing radius (R_0) of the Centre Surround receptive fields. Figure reported from figure 4 in [28].

The two Border Ownership responses are then combined together as follow.

$$\mathrm{B}^k_\theta(x,y) = \mathrm{B}^k_{\theta,L}(x,y) + \mathrm{B}^k_{\theta.D}(x,y)$$

For each pixel we have a response for all the angles evaluated for the circular integration along Von Minses distribution inspired annular receptive field (ν_θ) (figure 3.7), for each angle θ the two group of border ownership cells coding for opposite sides (θ and $\theta + \pi$) compete. Therefore for each pixel situation can be summarized as in figure 3.8. The winning border for each pixel is selected to be $\hat{\theta}$ if:

$$\hat{\theta} = \mathrm{argmax}_\theta \left(\mathrm{B}^k_\theta(x,y) - \mathrm{B}^k_{\theta+\pi}(x,y)\right)$$

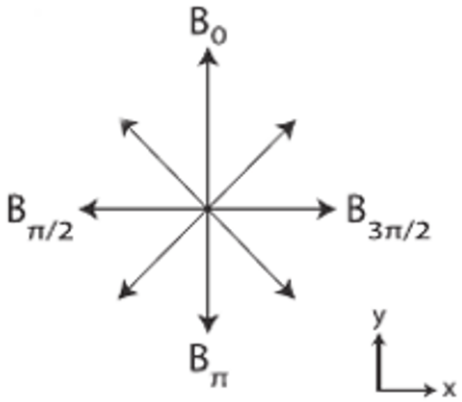


Figure 3.8: Border Ownership cells activity at each pixel. The magnitude of the activity is proportional to arrow length. Reported from figure 4 in [28].

5. **Grouping Cell Responses:** similar to what happens in the previous step now are the *Grouping cell* responses to be evaluated through a circular integration on annular receptive field considering, for each winner orientation of

Borner Ownership cells, the inhibition from the *Border Ownership* cell coding for the opposite direction. Grouping cell activity is defined as follow.

$$\mathrm{G}^k(x,y) = \sum_\theta \delta\left(\mathrm{B}^k_\theta(x,y), \hat{B}^k\right) \mathrm{x}\left[\mathrm{B}^k_\theta(x,y) - \mathrm{wxB}^k_{\theta+\pi}(x,y)\right] * \nu_\theta(x,y)$$

where

$$\delta\left(\mathrm{B}^k_\theta(x,y), \hat{B}^k\right) = \begin{cases} 1; & \text{if } \mathrm{B}^k_\theta = \hat{B}^k, \text{with } \hat{B} = \text{winning cell} \\ 0; & \text{if Otherwise} \end{cases}$$

The *Grouping* pyramid is the output of the grouping mechanism that now have to be normalized, merged in one unique level and finally the outputs from single feature channel have to be summed up together.

3.1.2 Global Structure of the Algorithm

The feed forward grouping algorithm is included in a big structure that is based on the one proposed by Itti et al. at the time when the first computational models of visual saliency came out. As preliminary step, from the input RGB Image information about single image features are extracted. This step is accomplished in order to see if any pop-out effect can make an area, or an object in the image, more quickly accessible to the visual attention.

At the end of the grouping mechanism a normalization operation is added in order to promoting groping activity of maps with few proto-object and suppressing the one with multiple proto-objects.

Feature Channels Extraction

The features chosen for the analysis are 3 but the total number of channels is 9 because two of them include more then one channel. The image is first of all divided in three color channels that come from the three matrix that compose the RGB image given as input: r, g, b. These channels are not directly used in the algorithm but are needed to extract 4 of the 9 total channels. Below are listed and briefly explained the main steps to extract the inputs for the single channels starting from the input RGB image.

1. **Color channels Extraction:** blue (b), red (r) and green (g) channels separated from the RGB image;

2. **Intensity channel Generation:** $I = \frac{r+g+b}{3}$, this intensity Image is than given to the grouping algorithm;

3. **Color channels Normalization:** each pixel intensity value in r, g, b smaller than the 10% of the maximum intensity value of the image are settled to zero. This can be done without altering algorithm performances because at very low luminance variations are not perceived;

4. **Broadly tuned color channels Extraction:** $red \rightarrow R = r - \frac{g+b}{2}$, $blue \rightarrow B = b - \frac{g+r}{2}$, $green \rightarrow G = g - \frac{r+b}{2}$, $yellow \rightarrow Y = \frac{g+r}{2} - \frac{|r-g|}{2} - b$;

5. **Half wave Rectification:** all the negative values in R, G, B and Y are settle to zero;

6. **Color opponency channels Generation:** $RG = R - G$, $GR = G - R$, $BY = B - Y$, $YB = Y - B$, these signals are then given to the grouping algorithm;

7. **Orientation channels Generation:** the four orientation channels O_α with $\alpha \in 0, \pi/4, \pi/2, \pi$ are created using I as input for the grouping mechanism and replacing the center surround mechanism with even Gabor filters with α orientations. The spatial frequency is chosen so that the width of the central lobe of the filters matches the zero crossing of the center surround original mechanism. The result of that is a center surround mechanism modulated by the orientation of the proto-objects. The specific Gabor filters are:

$$\mathrm{cs}_{on}(x, y) = \mathrm{e}^{-\frac{x'^2+\gamma^2+y'^2}{2\sigma^2}} \cos(\omega' x')$$

$$\mathrm{cs}_{off}(x, y) = -\mathrm{e}^{-\frac{x'^2+\gamma^2+y'^2}{2\sigma^2}} \cos(\omega' x')$$

where: $x' = xcos(\alpha) + ysin(\alpha)$ and $y' = ycos(\alpha) - xsin(\alpha)$ These operations are repeated changing α angle so 4 different channels are generated, all the result images are given to the grouping algorithm as input.

3.1.3 Normalization Step

At the end of the groping algorithm we have proto-objects conspicuities maps for all the channels. Every grouping pyramid has to be collapse to one single level. Each level is normalized and then a cross scale addition is accomplished. Following expressions show how conspicuity maps ($\bar{I}$, $\bar{C}$, $\bar{O}$) are generated, operator N_2 indicates the normalization process, while $\oplus$ indicates the cross scale addition achieved collapsing pyramid levels to a common scale. The scale to which collapse the pyramids is generally choose close to the middle scale. Russel et al. chose the eighth level to evaluate the model. For the each feature channel, conspicuity map is calculated as follow:

$$\bar{I} = \oplus_{k=1}^{k=10} N_2\left(G_I^k\right)$$

$$\bar{C} = \oplus_{k=1}^{k=10} \left(N_2\left(G_{RG}^k\right) + N_2\left(G_{GR}^k\right) + N_2\left(G_{BY}^k\right) + N_2\left(G_{YB}^k\right)\right)$$

$$\bar{O} = \sum_{\alpha \in 0,45°,90°,180°} N_2\left(\oplus_{k=1}^{k=10} N_2\left(O_\alpha^k\right)\right) \tag{3.1}$$

Conspicuity maps are then normalized again and linearly combined to form the proto-object based saliency map:

$$\mathrm{S} = \frac{1}{3}\left(N_2\left(\bar{I}\right) + N_2\left(\bar{C}\right) + N_2\left(\bar{O}\right)\right) \tag{3.2}$$

Once the fundamental structure of the model has been studied and understood, we approach the description of the extension made to the algorithm. The next sections describe how we decided to include information on depth discontinuity, the heart of the work was to understand how to provide information to the grouping mechanism so that it could recognize objects located on the same stereoscopic plane (figure 2.12).

3.2 Disparity Channel

The basic structure of the algorithm proposed by Russel et al. [28] and described in the previous section of this chapter have been left unchanged but a new channel is added to include three dimensional information. We refer to this channel as *Disparity Channel* and this section is entirely dedicated ti its description. The model, as the one proposed by Russel et. al, is implemented using MATLAB (Mathwork, Natik, MA, USA). As all the other channels in the model, the *Disparity Channel* receives an input, two RGB stereo images of the same scene, and extracts information to give to the grouping algorithm. In this first version of the model an other input is also required. The following list gives a generic introduction of the inputs that the *Disparity channel* requires and the next two sections are dedicated to their more detail explanation.

Inputs required by the new channel are the following.

- **A couple of stereo images:** the stereo images have to be rectified.
- **A set of expected values of stereoscopic disparities:** disparity values are provided as numerical values indicating the horizontal displacement between the position of two corresponding points in the two stereo images. The value is given in terms of number of pixels along the horizontal line and its sign is an indication of the verse of the shift.

From the input images, taking advantage of information about the range of disparities, a series of steps is made in order to create the right images on which the grouping algorithm can act so that this can be functional in order to search depth discontinuities in the proposed visual scene. These steps are listed and summarily described below.

1. **Contrast Evaluation:** the two stereo images are divided in square blocks, for each block the *Root Mean Square contrast* is calculated and compared to a threshold values, if *Root Mean Square contrast* of both stereo images, is below the threshold that block is considered a low contrast zone and is excluded from saliency computation.

2. **Cost Function Calculation:** at each pixel for each disparity value the corespondent *Sum Of Absolute Difference* is evaluated between the two stereo-images (see section 3.2.4). This step is repeated for each image pixel and at each disparity values in the chosen set. When at the current disparity there is a good match between the two stereo images the result of the cost function is low, while it is high if there is no match between the pixels from the two images and at the current disparity.

3. **Maps Normalizaiton:** after the previous step, for each disparity value, is created a map that have the same dimensions of the input images and where pixels values are the results of the Cost Function Calculation. In these maps items situated at the current disparity have low pixels intensity values then a series of normalization operations are carried out to reverse the trend: we want good matches between the stereo images to be marked by high values in the correspondent area of the maps.

4. **Disparity Opponency Evaluation:** each maps (one at each disparity) is now subtracted from all the other maps.

The scheme in figure 3.9 shows the steps required to get the right information to give to the grouping algorithm.

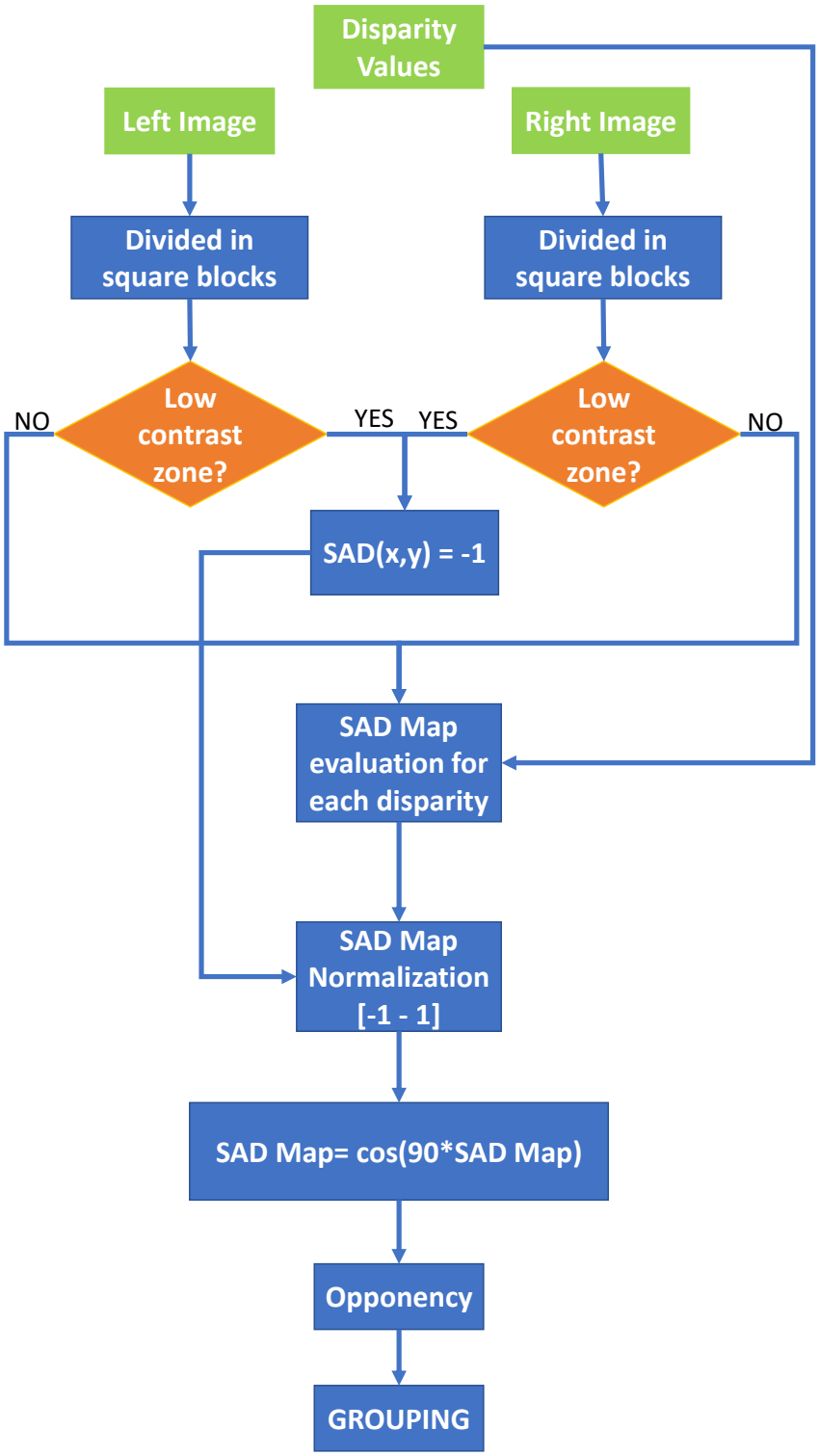


Figure 3.9: Overview of the Disparity Channel Algorithm: from the inputs to the start of the grouping mechanism.

3.2.1 Input Images

Is it critically important that the two stereo images given to the *Disparity Channel* are rectified.

Image Rectification

The rectification of an image can be interpreted as a prospective correction, it is a transformation process generally used to project more images on a common two-dimensional surface. It is also used to correct single images to a standard coordinate system (eg: a rectangular, Cartesian coordinate system). In figure 3.10 it is shown how different possible perspectives can memorize visual information about the same object and then, after a rectification process, the two "points of view" have the same projections system. This permits more accurate comparison analysis [5].

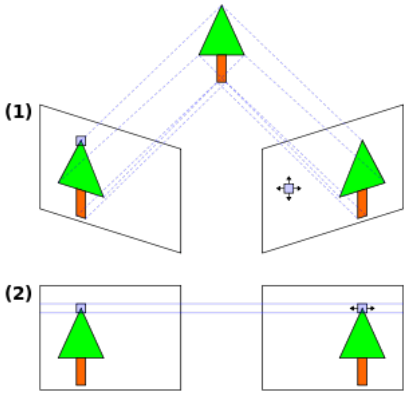
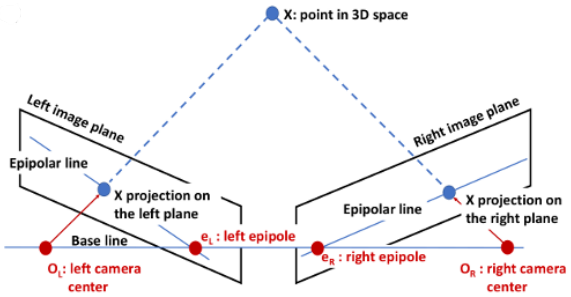


Figure 3.10: Image Rectification. **(1)** Two stereo images before rectification and relative search space. **(2)** Same two stereo images after rectification and relative search space, the images now have the same projection system. Reported from [42].

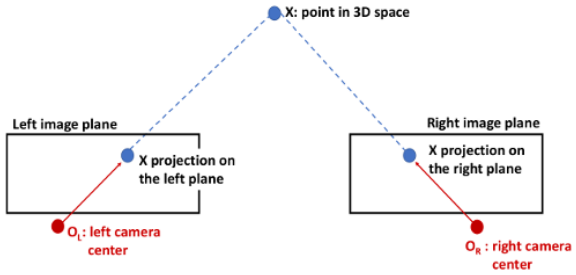
A rectification system is often included in the camera used to shoot stereo images but sometimes manual rectification may be required and it can be implemented following widely documented algorithms. MATLAB has some functions that can rectify images by providing some parameters of the camera used.

Our algorithm works under the hypothesis that the two stereo images that are provided as inputs, have been already rectified in order to simplify the search of matching points, required for establishing the correct disparity between the positions of a specific point in the two different views. In fact the algorithm uses *Binocular disparity* between the two stereo images to estimate the distance of a specific area or object in the scene from the observer. This means that it looks for the displacement of a specific point between the two stereo images and if the images are rectified, the vertical coordinates of corresponding points are identical so the search for correspondences can be carried out exclusively along horizontal lines. At the contrary

if the images were not rectified, the search for two corresponding points should be carried out along two dimensions. This would greatly complicate the algorithm and increase the computational cost as well as reduce execution speed. We choose not to implement a rectification mechanism in the algorithm itself because any of them requires to know specific parameters of the camera such as the location of the camera in the 3-D scene (extrinsic parameters), the optical center and focal length of the camera (intrisic parameters) and the lens distortion. These parameters are all needed to find a projective transformation with which rectify the images. Referring to *epipolar geometry*, that explains the rules to link the three-dimensional world to the two-dimensional representation in stereo vision, in two rectified stereo images the base line is parallel to the image planes and epipolar lines are not defined. For the scheme and the definition of epipolar geometry in the case of non rectified and rectified stereo images look at figure 3.11. One of the easiest way to shoot two rec-



(a) *Not rectified stereo images*



(b) *Rectified stereo images*

Figure 3.11: Epipolar Geometry: scheme and definition. **(a)** When the stereo images are not rectified the line that links the cameras center (base line) intersects the two image planes so epipolar lines can be defined. **(b)** When the stereo images are rectified the base line is parallel to the image planes so epipolar lines can not be defined.

tified stereo images, even if it is the less precise, it is to use a normal camera and and move it horizontally between the two image shots.

For what concerns analogies with the human biology, our eyes are two cameras that are with good approximation, located on the same horizontal axis. Anyway images from the eyes have to be projected in the in the retina that has a convex

shape and so they are distorted. In reality any mechanism similar to the rectification process is known to happen in human vision system but what we know is that only horizontal position shifts, between the view from left eye and the one from right eye, is the only considered to elaborate depth perception from stereopsis (for more detail information see section 2.2.1).

3.2.2 Disparity Values and Direction

As anticipated, a part from the two stereo images, the algorithm required an other input: a set of disparities expected to be found between the two stereo images. These values are given as a vector of integers that indicates the entity, expressed in pixels number, of the horizontal shift of the same point between the two stereo images. If we consider a certain pixel in the left image from a stereo pair, and we look for the correspondent one (the one that represents the same point of the visual scene) in the right image, the horizontal shift between these two correspondent point can both be positive (to the right) and negative (to the right). For the correct interpretation of the disparity values and their verses it is necessary to know which of the two views is taken as reference. If the algorithm must look for two matching points between the two stereo views, given the disparity values and their verses, the displacement absolute value does not change if the search is made from the right image to the or from the left image to the right image , but the verse (sign) of the disparities changes.Knowing a priori the range of disparities between two stereo images is not easy. In general this depends on the size and resolution of the images, the way in which the two cameras are positioned to take the two stereo images and the distortion introduced by the cameras themselves. Cameras with low distortion and in a perfectly parallel alignment take stereo rectified images where disparity values have all the same sign.

This first version of the algorithm do not include an automatic evaluation of the disparity values and for that reason a set of probable disparity values has to be provided as an input. This can be easily done if the images are artificially created so that the disparity values are chosen at the moment of the image creation creation (see section 4.1). For what concern testing the algorithm with natural scenes images as inputs, it is, for now, necessary to find or create a dataset that includes, as well as a series of pairs of rectified stereo images, also the relative vector of the horizontal disparities associated. Looking at datasets available online [16] it is possible to understand that a first general rule to choose the disparity range is that the maximum disparity is general less then one tenth of the image width itself. The idea is to contribute to the creation of a final computational stereo vision model that will be able to automatically chose the range and subsets of disparities to evaluate. It will follow the limit imposed by the resolution and the working method of far, near and tuned zero cells of the visual cortex that respectively respond to disparities

in planes further away from the fixation plane, closer than the fixation plane and on the fixation plane. These planes are all included in the *Panum's area* (see section 2.2.1 for further information about stereopsis and human biology). From this point we will use the expressions *disparity* and *horizontal shift* without any difference, to indicate one of these numeric values provided as inputs.

Now each pixel of the image has to be assigned to the correct value of disparity within the chosen range. For that reason a specific *Cost Function* (see section 3.2.4) is utilized. This function works looking for similar areas in the two stereo views but all the areas of the images where no features occur (i.e. uniform background) will perfectly match even if they are not placed in correspondent positions in the two stereo images, making the search less robust. Furthermore these areas are generally not considered a salient area. For that reason, we choose a priori to exclude low contrast areas. In the next section the way we used to evaluate the contrast is explained.

3.2.3 Contrast Evaluation

Low contrast zones are not considered salient areas and, as will be more clearer after the next section, letting these zones be included in the next steps would make the following results unreliable. Furthermore, excluding low-contrast areas reduces the total computational cost. To decide whenever an area of the visual scene is a low contrast areas or not the algorithm proceeds following these steps:

1. **Images Divided in blocks:** for two pixels situated at the same position in the two stereo images, one in the right view and one in the left view, a square block centered in the pixels themselves is considered.

2. **Root Mean Square Contrast Calculation:** for each couple of pixels, in the respective blocks, the *Root Mean Square Contrast (RMS)* is calculated as follow.

$$\mathrm{RMS}_{constrast} = \sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(\mathrm{I}_{ij} - \mathrm{I}_{av})^2} \tag{3.3}$$

   where:

   - $\mathbf{I}_{ij}$ is the i^{th} j^{th} element of the block with size MxN;
   - $\mathbf{I}_{av}$ is the average intensity of all the pixel values in the image.

3. **Comparison to a threshold value:** the value of $RMS_{contrast}$ (RMS_C) of each block is compared to a threshold value.

4. **Final Decision:** depending on the result of the comparison two decisions can be made:

- RMS_C of both stereo images $< Threshold \rightarrow block = low\ contrast\ zone$: the whole block is excluded from the following steps;
- RMS_C of both stereo images $> Threshold \rightarrow block \neq low\ contrast\ zone$: the whole block is included in the following steps.

3.2.4 Cost Function and Sum of Absolute Differences Map

The next steps of the algorithm included in the *Disparity channel* evaluation are the real core of the channel itself. The ultimate goal is to prepare the right inputs for the grouping mechanism, fully described in section 3.1.1. We want to make this mechanism to operate on specific depth planes (see figure 2.12 for reference). Items lying on a common depth plane are characterized by similar values of horizontal shift (disparity) between right and left view. Therefore to simulate the Gestalt Principles (section 2.3.3 and [21] for reference) of continuity and proximity (figure 2.20) to identify the presence of a proto-objects structure in a specific depth plane, grouping mechanism has to get as inputs images that recreate single depth planes. A depth plane is associated with a specific value of horizontal shift so we have to divide the scene grouping together all the areas characterized by the same values of this shift. The horizontal shifts that are possible to find between the two view are given, in term of pixels numbers, as input, so the algorithm works calculating the result of a specific cost function which indicates, for each value of disparity, the goodness of the matching between two areas: one considered in the right view of the stereoscopic pair and the other in the left view. The cost function is the *Sum of Absolute Distances (SAD)* that is a measure of similarity between image blocks widely used in stereo matching problems. The **Sum of Absolute Distaces (SAD)** between two image blocks with the same dimensions is calculated as follow:

1. difference between each corresponding pair of pixels: one for each block $\rightarrow$ new difference block with the same size (*Difference*);
2. calculation of the absolute value of the whole difference block (*Absolute*);
3. pixel values within the absolute difference block are adding together (*Sum*).

The more the two blocks are similar, the more the result of the *Sum of Absolute Differences* is low.

For what concerns the *Disparity channel* algorithm these three operations (SAD) are of critically important for the creation of maps that we call SAD maps whose basic steps are listed below. The correct processing of these maps is essential for the success of the simulations that the entire model complete. *SAD maps* generation process steps are the following, remember that all of these have to be repeated for each disparity (horizontal shift) value given as input:

1. a disparity (horizontal shift) value is picked from the set;
2. a pixel in one of the two stereo images (suppose the left one) is selected to be the center of a squared block in the current view;
3. in the other image (right view) the pixel located in the same horizontal line, as the chosen pixel in the other view (left view), but shifted by the value indicated by the current disparity, becomes the center of a squared block in the right view → the two blocks respectively in the left and in the right view have to be of the same size;
4. Sum of Absolute Difference is evaluated between the two image blocks;
5. a SAD map for the current disparity is created setting the map pixel, in the position corresponding to the one that the pixel under analysis has in the original image (left view), to the value of the Sum of Absolute Differences just calculated.

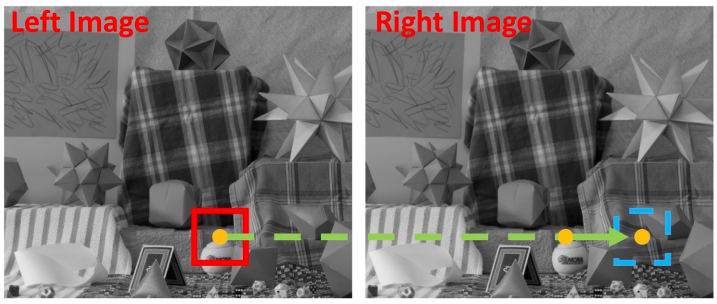
These steps have to be repeated for each pixels of the images, excluding those belonging to a low contrast area, and for each disparity in the set. Figure 3.12 shows how the value is assigned to a pixel in a *SAD map* for a specific disparity value. In the next small subsections the following information are given:

- **SAD map definition**;
- **low-contrast areas management**;
- **maps preparation for the next steps of the algorithm**.

SAD Map

The map we called *SAD map* in reality is an image that has the same dimensions of the two stereo images given as inputs and where each pixel value corresponds to the *Sum of Absolute Differences* result for that specific pixel for one specific disparity. Therefore at the end we have as many SAD maps as the disparity number in the set. Each pixel in a *SAD map* can have a value that goes from 0 to a maximum values that varies with the image scale and the dimensions of the blocks used to calculate the *Sum of Absolute Differences*. Given a disparity value of the input set and assuming not to have developed an alternative algorithm for managing low-contrast areas, pixel values in the map give the following information depending on their value.

- **High pixel value in the map:** indicates that the two blocks from the two stereo images do not have a good match at the current disparity;



(a) *Stereo Images*

$$abs\left(\begin{matrix}103 & 86 & 204\\ 99 & 92 & 205\\ 101 & 119 & 105\end{matrix} - \begin{matrix}104 & 111 & 229\\ 99 & 109 & 120\\ 145 & 112 & 113\end{matrix}\right) = \begin{matrix}1 & 25 & 25\\ 0 & 17 & 85\\ 44 & 7 & 8\end{matrix}$$

SAD= 212

(b) *Example of SAD numeric calculation*

Figure 3.12: Sum of Absolute Differences: **(a)** Given a certain disparity value, for each pixel of one of the stereo images (Left Image here) a squared block is considered (red contours), in the other stereo image (Right Image here) the pixel taken as the center of a block of the same size (dashed light blue contours), is situated in the position moved by the disparity under analysis, along the horizontal line, from the one of the pixel in the Left Image;**(b)** Sum of absolute values of the differences between the right and the left block.

- **Low pixel value in the map:** a value equal or close to zero indicates that one of the two situations listed below occurred.

  1. the two blocks from the two stereo images have a good match at the current disparity;
  2. the two blocks individuate a low contrast zone. These two cases have to be distinguished and for that reason since the beginning the low contrast areas have a different treatment.

Low contrast areas management

At the end of the contrast evaluation process (section 3.2.3) the algorithm already has the information about low contrast zones locations. There fore before starting the calculation of each SAD map, the pixel values at the low contrast area are preset to a value of -1 in the map. So we know that the values of the SAD map can be interpreted as follow:

- **High pixel value in the map:** indicates that the two blocks from the two stereo images do not have a good match at the current disparity;

- **Low pixel value in the map (zero or close to zero):** indicates that the two blocks from the two stereo images have a good match at the current disparity;
- **Pixel value equal to -1:** the two blocks individuate a low contrast zone.

In this way, data interpretation problems are avoided because they try to exclude ambiguous situations.

Normalization and Cosine Operator

To facilitate next steps we want to mark a good match, at a specific disparity with a high value in its location on the SAD map. For that purpose, values of the map are:

1. normalized between -1 and 1 $\rightarrow SAD_{norm}$;
2. replaced according to the following operation: $SAD_{map} = \cos(90 * SAD_{norm})$.

Now for each disparity of the chosen set we have a map where:

- **Null pixel values:** indicate a low-contrast zone or a zone that is not located at the current disparity;
- **Low pixel values (close to zero):** indicate a zone that is not located at the current disparity but that is closer than the ones characterized by null values;
- **Pixel values equal or close to 1:** indicate a zone that is located at the current disparity.

Example case: from input images to the final SAD map

In order to better understand the operation of the disparity channel algorithm,an example case is studied, from the input images to the SAD_{map} obtained after the cosine operator. The steps are shown in figure 3.13 and the following explains what the individual panels in figure 3.13 represent.

1. **Input Stereo Images:** first raw in figure 3.13 shows left view and right view given to the algorithm as inputs. The couple of stereo images is composed by 2 images of 7 gaussian blobs of random dots on an uniform background. Between two correspondent blobs (one in the left and one in the right view) a known horizontal shift has been applied.

   For example in the case in the figure:

   - 3 blobs at the left of the scene are at disparity +5: shifted of 5 pixels to the right from the left to the right view;

- 3 blobs at the right of the scene are at disparity -5: shifted of 5 pixels to the left from the left to the right view ;
- 1 blob in the middle of the scene is at disparity 0: it has not been shifted, same position in the left and in the right view.

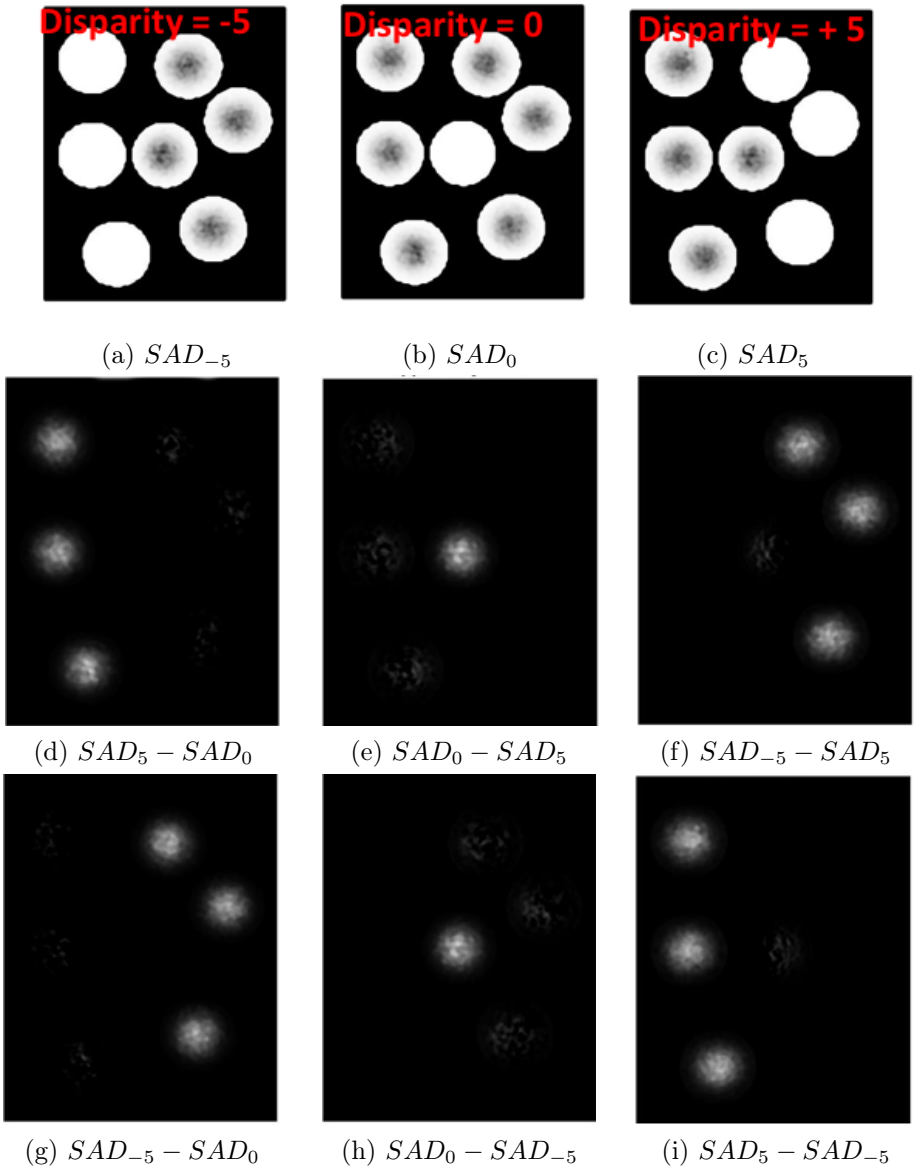
2. SAD_{map} **after Normalization:** the second row in figure 3.13 shows the three SAD maps evaluated for one of the three values of horizontal shift $(-5, 0, +5)$ and after the normalization in the range [-1 − 1]. In these SAD_{map}s blobs at the current disparity have null pixel values (green in figure 3.13), low contrast zones (the whole background) is settled to -1 (blue in figure 3.13) and blobs that are not at the current disparity show highest SAD values (from yellow to red pixels in figure 3.13).

3. SAD_{map} **after cosine operation:** the third row in figure 3.13 shows the three SAD maps obtained after the cosine operation. In these SAD_{map}s blobs at the current disparity have pixel values equal to 1 (red in figure 3.13), low contrast zones have null pixel values (blue in figure 3.13) and blobs that are not at the current disparity show lowest SAD values 3.13).

After the cosine operation a sort af opponency between all the disparities is recreated following the scheme of Color Opponency channel in Russel et. al model [28], this is the last step before letting the Grouping algorithmrun.

3.2.5 Disparity Opponency

The last step before giving depth information to the grouping algorithm, aims to recreate the different depth planes corresponding to different values of disparity. We want to simulate *Grouping Cells* activity on each depth planes in order to see if recognizing proto-object structures at different distances from the observer and use them to search for depth discontinuities in the scene somehow modify and hopefully improve the prediction of fixation points compared to the two-dimensional one. The grouping algorithm is able to recognized objects if the pixels intensity representing them, make it possible to use the Gestalt psychology principles of proximity and continuity ([21] and figure 2.20 for references). For simplicity, for each disparity, and so for the respective depth plane as well, we chose to represent with the highest pixel values only the area located at the current disparity (depth) so that they can be recognized as perceptual object structures. To be sure that map given as input to the grouping algorithm shows the highest values only in areas located at the current disparity, each SAD_{Map} is subtracted from all the other, akin to the Color Opponency Channel in Russel et al. model [28]. For example, if there are 3 disparity values (disparity -5, disparity 0, disparity +5), as in figure 3.14, we fist evaluate 3 SAD_{Map}s (one for each disparity, first row in figure 3.14) and subtractions yeld 6 new map:



(a) SAD_{-5} (b) SAD_0 (c) SAD_5

(d) $SAD_5 - SAD_0$ (e) $SAD_0 - SAD_5$ (f) $SAD_{-5} - SAD_5$

(g) $SAD_{-5} - SAD_0$ (h) $SAD_0 - SAD_{-5}$ (i) $SAD_5 - SAD_{-5}$

Figure 3.14: Disparity Opponency. Results from the disparity opponency subtractions after the half wave rectification. All the 6 opponency maps (second and third row) are obtained starting from a couple of artificially created stereo images where the set of disparity is $-5, 0, 5$. First row (white pixels have the highest intensity values): **(a)** SAD_{map} after cosine operation at disparity -5, **(b)** SAD_{map} after cosine operation at disparity 0, **(b)** SAD_{map} after cosine operation at disparity -5. Second and third rows show the results of the subtractions, after the half wave rectification: these maps are all given to the grouping algorithm as inputs. For each disparity there are two new maps representing the respective depth plane.

$$\text{Map}_{5,0} = \text{SAD}_5 - \text{SAD}_0; \text{Map}_{0,5} = \text{SAD}_0 - \text{SAD}_5;$$
$$\text{Map}_{-5,5} = \text{SAD}_{-5} - \text{SAD}_5; \text{Map}_{5,-5} = \text{SAD}_5 - \text{SAD}_{-5};$$
$$\text{Map}_{-5,0} = \text{SAD}_{-5} - \text{SAD}_0; \text{Map}_{0,-5} = \text{SAD}_0 - \text{SAD}_{-5};$$

In general with a set of N disparities the Opponency Maps to give to the grouping algorithm are

$$(N) * (N-1) \tag{3.4}$$

We already know from the description of the grouping mechanism (see figure 3.4 and section 3.1.1) that the actual inputs for the real core of the grouping algorithm are given by the result of the half-wave rectification of the maps resulting from the opponency subtractions (figure 3.14 shows the results of the opponency differences after the half-wave rectification).

The following two sections describe the steps that are performed by the model after the groping mechanism has been applied: first the conspicuity from the *Disparity Channel* has to be elaborated from the resulting grouping pyramid and then it has to be fused with the map resulting from the two-dimensional model (composed by intensity, color opponency and orientation channels which are fully described in section 3.1).

3.2.6 Normalization and Merge of Pyramid Levels

At the end of the grouping algorithm, for what concern the *Disparity Channel* results we have an images pyramid for each Opponency map. In order to obtain a single conspicuity map ($\bar{D}$) to add to the one coming from the other two-dimensional channels the pyramids must be normalized level by level and merged to a common scale that generally is closed to the middle level of the original image pyramid. The normalization steps are the same used for the *Orientation Channel* and briefly shown in equation 3.1. So for the *Disparity Channel* we have:

$$\bar{D} = \sum_{OpponencyMaps} N_2 \left(\oplus_{k=minlevel}^{k=maxlevel} N_2 \left(D^k_{SingleMap} \right) \right) \tag{3.5}$$

Where the N_2 operator indicate a series of operation the are left uncharged from the Itti et al. model dated 1998 [15]. Each level is normalized by following the steps listed below.

1. **Maps normalization in a fixed range:** suppose that the chosen range has minimum value equal to 0 and maximum equal to M ($[0.....M]$).

2. **Threshold setting:** $threshold = min(range) + \frac{max(range) - min(range)}{10}$.

3. **Local Maxima Search:** considering the original map, excluding the first and last rows of pixels and the first and last columns of pixels as reference see figure 3.15.

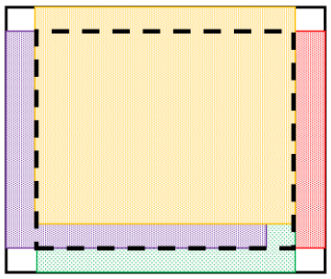


Figure 3.15: Image blocks for normalization process. The square with continuous black contours is the entire image; the square with dashed black contours is the reference block in which pixels are chosen, their value is compared to the corresponding (a the same location) pixel value in all the others blocks: orange block is moved one row higher, green block is moved one row lower, red block is moved one column to the right and purple block is moved one column to the left.

The reference image block is analyzed pixel by pixel and if all the following condition are satisfied the current pixel is memorized as a local maximum. Conditions to be satisfied are:

- current pixel value > value of the pixel at the same location in the image block moved one pixels row higher (orange block in figure 3.15);
- current pixel value > value of the pixel at the same location in the image block moved one pixels row lower (green block in figure 3.15);
- current pixel value > value of the pixel at the same location in the image block moved one pixels column to the right (red block in figure 3.15);
- current pixel value > value of the pixel at the same location in the image block moved one pixels column to the left (purple block in figure 3.15).
- current pixel value > $threshold$.

4. **Local maxima Information Extraction:**

- total number of local maxima;
- average of local maximum values;
- sum of all local maximum values.

5. **Map Normalization:**

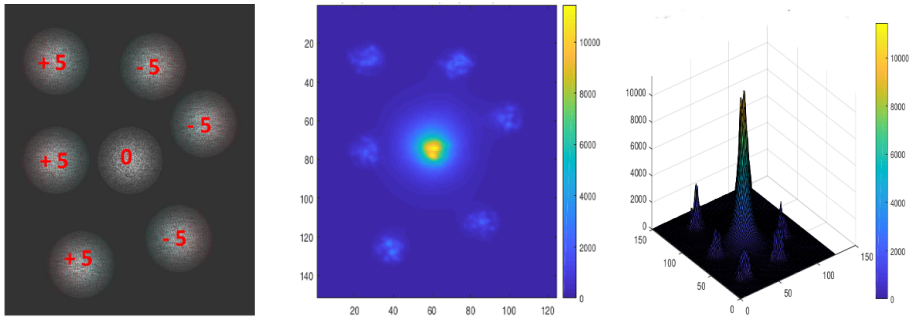
$$\text{Map}_{norm} = \begin{cases} (M - \bar{m})\text{Map}; & \text{if } n^{\circ}\text{local maxima} > 1 \\ M^2\text{Map}; & \text{if } n^{\circ}\text{local maxima} = 1 \\ \text{Map}; & \text{if } n^{\circ}\text{local maxima} = 0 \end{cases}$$

In the previous expression Map_{norm} is the result of the normalization, M is the output of the gropung mechanism and $\bar{m}$ is the average of the local maxima values.

This normalization allows to weigh the individual maps, giving emphasis to the possible pop-out effect. In fact, the weight increases as the difference between the value of the global maximum and the average value of the local maxima increases and it is easy to understand that if a feature map is affected by the pop-out effect, the peak intensity value will be much higher than the average value of the local maxima. All these operations are repeated for each level of each pyramid.

Than the levels are scaled to a common scale, using a bicubic interpolation (read the introduction to section 3.1 for further information). The scale to which scale all the levels can be selected time by time the model runs but generally if it is close to the middle level of the images pyramid, no important information from the other levels should be lost. After that the normalization is repeated again the last step is adding the resulting maps (one for each considered Opponency Channel).

Figure 3.16 shows the resulting $\bar{D}$ map obtained running the entire disparity channel algorithm, given as input the pair of stereo images taken into account in the case shown in the figure 3.13. As can be seen in figure 3.16, the algorithm is able to detect the presence of the only object located at zero disparity that, with all the other features being equal, is expected to be the most salient object.



(a) *Input Images Red-cyan anaglyph* (b) *Disparity Saliency Map: 2D representation* (c) *Disparity Saliency Map: 3D representation*

Figure 3.16: Disparity Channel Output. **(a)** Red-cyan anaglyph from the two stereo images given as input to the disparity channel: red numbers indicate the horizontal shift relative to each blob and expressed as pixels number. **(b)** 2D representation of the Disparity channel Saliency Map obtained as output from the disparity channel algorithm: high pixels value (yellow) indicate the presence of a salient zone. **(c)** 3D representation of the Disparity Channel Saliency Map obtained as output from the disparity channel algorithm

3.3 Sum of Two-dimensional and Disparity Saliency Maps

Grouping algorithm is applied simultaneously to all the feature channel in figure 3.1. Therefore, the last step is to combine the two-dimensional map (includes contribution of intensity, color opponency and orientation channels) with the contribution of the disparity channel that add depth information to the whole model. The two map are weight and summed together (linearly combined together) as follow:

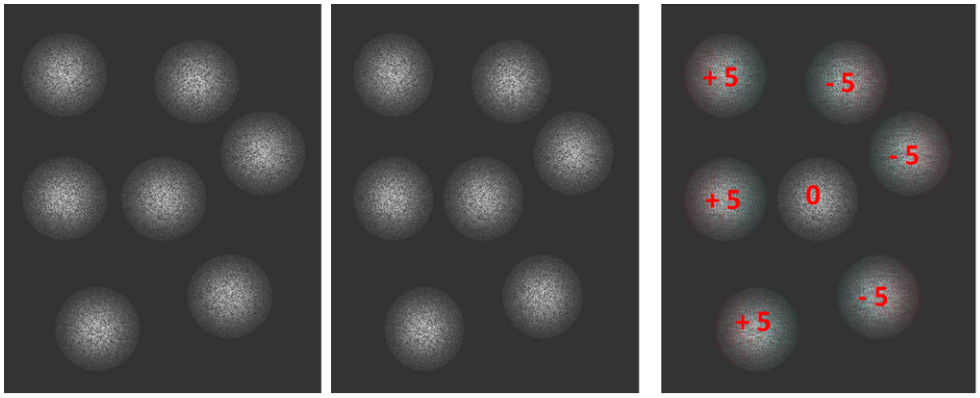
$$\text{SaliencyMap}_{final} = w_d \bar{D} + w_s \text{S}$$

where:

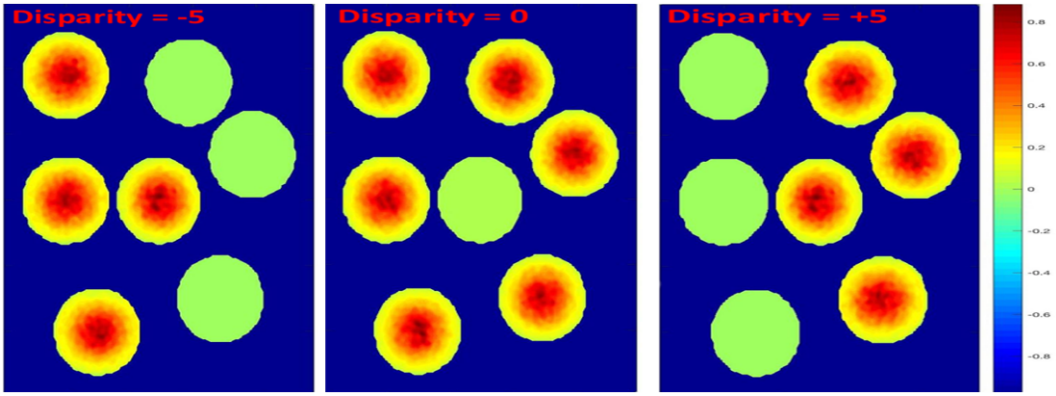
- $\text{SaliencyMap}_{final}$ is the final saliency map that accounts both disparity channel information and two-dimensional information from intensity, color opponency and orientation channel;
- $\bar{D}$ is the saliency map that accounts only the disparity channel contribution;
- S is the saliency map that accounts only fro two-dimensional information from the intensity, color opponency and orientation channels (see equation 3.3)
- w_d multiplicative factor on which depends the weight of the results obtained from the disparity channel in the final map;

- w_s multiplicative factor on which depends the weight of the results obtained from the intensity, color opponency and orientation channels, considered together as a unique two-dimensional saliency map, in the final map.

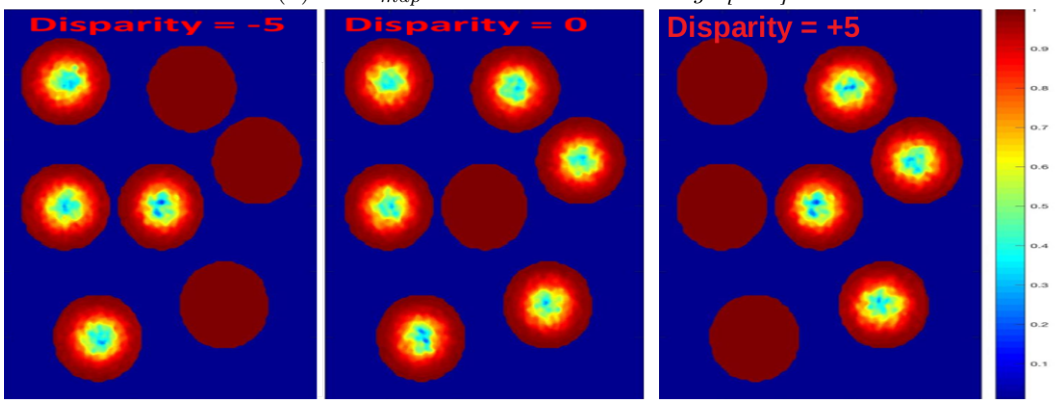
The choice of the two weights does not follow a precise rule or specific knowledge about human biology. However, common sense suggests that we can not consider a 1: 1 ratio but rather we believe that the map that takes into account the three two-dimensional features must weigh more than the added contribution from the information to the binocular disparities added by the disparity channel. For that reason, even if the weights can be settled each time the algorithm runs we almost always set a weight equal to 20% for the disparity channel contribution ad a weight equal to 80% for the two-dimensional contribution. This makes depth information comparable with information from single two-dimensional feature channels but substantially smaller than the whole two-dimensional contribution that comes from the global set of the three different feature channels.



(a) *Input Images and Red-cyan anaglyph*



(b) *SAD_{map}s normalized in the range [-1-1]*



(c) *SAD_{map}s after the trend inversion through the cosine operation*

Figure 3.13: Example case: from input images to the final SAD map. **(a)** from left to right: left view, right view and Red-cyan anaglyph. Red numbers are blobs disparity values: given as input to the disparity channel algorithm (-5,0,+5).**(b)** from the left to the right: SAD_{map} at disparity -5, SAD_{map} at disparity 0, SAD_{map} at disparity +5. The maps are normalized between -1 (blue) and +1 (red). **(c)** from the left to the right: SAD_{map} at disparity -5, SAD_{map} at disparity 0, SAD_{map} at disparity +5. The trend has been reversed by a cosine operation, maps are normalized between 0 (blue) and +1 (red).

Chapter 4

Results

Between the previous chapter, entirely dedicated to the model description, and the next one that explains the evaluation techniques used to quantify model performances, it was decided to insert this short intermediate chapter describing the experiments carried out to test the correct operation of the model itself. The main purpose is to discover if the extension of the existing proto-object saliency model [28] [27] introduces any changes and and hopefully improvements to the model results. In order to discover that, it is important to follow the steps listed below.

1. **Evaluate that the added channel works as it is supposed to:** we artificially created couples of stereo images where the only feature contributing to potentially grab human attention is the different distribution of the items across depth planes (figure 2.12) These tests only evaluate the correct contribution of the disparity channel so the two-dimensional part of the algorithm does not run for the first experiments.

2. **Compare results obtained from the two-dimensional model alone with those obtained by adding the disparity channel:** for that purpose is needed a dataset that includes the following data.

    - **Set of rectified stereo images pairs:** to use as inputs for the model.
    - **Information about stereo disparity values:** to use as input for the disparity channel.
    - **Reliable ground truth:** to test the performances quality of the model.

As anticipated in chapter 2, two are the common outputs for a generic visual selective attention models (see figure 2.13):a saliency map and the focus of attention track extracted from the saliency map itself. Since our model has saliency map as its sole output, its evaluation consists in observing the similarity of its outputs maps with the maps provided as ground truth by any online dataset. Frequently, as in the case of the dataset used, the ground truth is obtained through eye tracking experiments.

In the case of the dataset we used, eye-tracker results from 35 different observes have been post-processed into fixation density maps. Chapter 5 is entirely dedicated to the evaluation of the model through specific statistic comparisons with the fixation density maps, in this chapter results are shown only through visual comparison. The paucity of datasets with both stereoscopic views and a valid ground truth makes the validation of this approach particularly difficult. This chapter gives a briefly description of both the creation of the artificially images to test the disparity channel performances and the dataset used to test any changes made to the existing model. For all the tests described in the following chapter, both for artificially stereo pairs (section 4.1) and real scene images (section 4.2), model parameters are listed in table 4.1.

Parameter	Meaning	Value
levels number	number of levels in the image pyramid	10
collapse level	normalization collapse level	8
γ	edges detector Gabor filters parameter	0.5
σ	edges detector Gabor filters parameter	2.24
ω	edges detector Gabor filters parameter	1.57
σ_i	center surround inner standard deviation	0.9
σ_o	center surround outer standard deviation	2.7
$w_o pp$	inhibitory signal weight (center surround)	1
R_0	center surround zero cross radius	2
w_b	inhibitory signal weight (border ownership cells)	1
σ_1	orientation channel Gabor filters parameter	3.2
γ_1	orientation channel Gabor filters parameter	0.8
ω_1	orientation channel Gabor filters parameter	0.7
M	maximum values for normalization final step	10
neighborhood size	contrast evaluation (disparity channel)	7x7
threshold	contrast evaluation (disparity channel)	0.01
neighborhood size	SAD maps evaluation (disparity channel)	7x7

Table 4.1: Model parameters

4.1 Artificial Stereo Images Pairs: Disparity Channel Results

In order to see if the disparity channel algorithm is able to detect disparity discontinuities between two stereo views of the same scene, we need to have couples of simple stereo images where the only relevant feature to find is the item (or the

items) that lies in depth plane different from the ones where the other items are. Figure 4.1 aims to remember what we mean with the concept of depth plane.

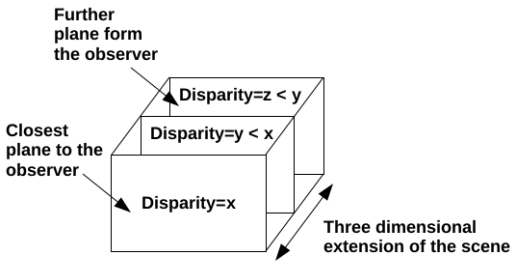


Figure 4.1: Disparity or depth planes. Three dimensional scene is decomposed in a series of planes parallel to the coronal head plane of the observer: each plane is associated to a stereoscopic disparity that is expressed in the form of horizontal shift pixel numbers (x,y,z) and that decreases with increasing distance of the plane from the observer.

4.1.1 Stereo Images Pairs Creation

To test the function of the disparity algorithm, we artificially created couples of stereo images. Each couple is composed of two images of 7 gaussian blobs of random dots on a uniform background. Between two views of the same image a horizontal shifts is applied to all the Gaussian blobs, to see if the algorithm can detect the presence of anomalous disparities. The composition of these images was performed using the MATLAB2017b software in order to recreate pairs of rectified stereo images. Setting precise disparity values, each one associated with specific items in the image, allows us to know a priori their relative distances from the observer. This let us know, according to the fundamental hypotheses on which our model is based, where is the most salient item that the algorithm should detect. Each stereo images couple is composed by two view that we will call *right view* and *left view* by analogy with the projections of the right and left retina.

Left view: created only once and left unchanged for all the images couples.

Right view: created specifically for each images couples on the basis of the disparities that we want to associate with each item in the figure.

Artificially Created Left View

The left view of all the artificially created stereo couples is build as explained below.

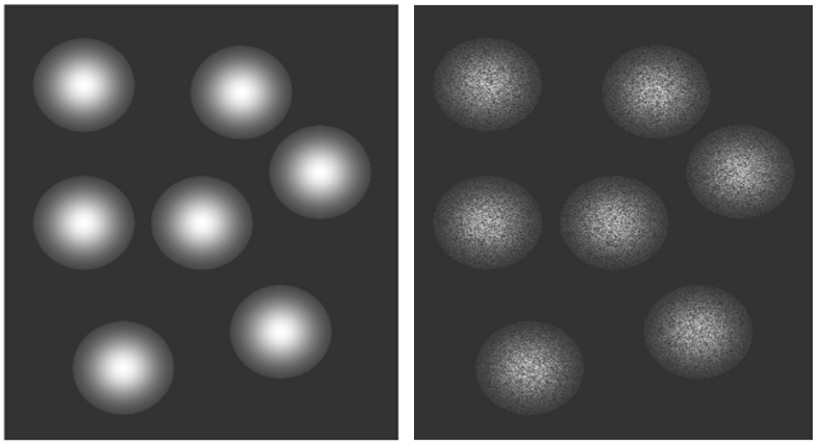
1. **Creation of a two-dimensional space:** a matrix of dimensions 601x501.

2. **Gaussian function positioning:** 7 identical two-dimensional Gaussians are created and centered at different points of the created grid.

3. **Gaussian truncation and Gaussian Blobs creation:** for every Gaussian function all the values below 10% of the maximum are forced to be 0. This makes sure that there is no overlap between functions and thus allowing to isolate 7 Gaussian blobs (figure 4.2**(a)**).

4. **Random dots insertion:** the grid of Gaussian blobs is not use as it is to evaluate the correct functioning of the algorithm. This choice is made to avoid the perfect central symmetry (rare to find in natural scene images) of the Gaussian functions to interact with the algorithm mechanisms. For that reason instead of using a composition of simply Gaussian blobs random dots are added: a grid of random dots is created (with the same dimensions of the original one) and then is filtered using the mask composed by the distribution of the 7 Gaussian blobs obtained at the end of step 2 and shown in figure 4.2**(a)**. The resulting image is shown in figure 4.2**(b)**.

   The left view has to be considered as the reference one for every test: when a blob "is moved to the right (left)" it means that its position in the right image changes along the horizontal line, of a given number of pixels to the right (left) relative to the position of the corresponding blob in the left view.



(a) *Artificial stereo view before random dots insertion*

(b) *Artificial stereo view after random dots insertion*

Figure 4.2: Artificially created stereo left view. **(a)** figure shows the 7 Gaussian blobs before the random dots insertion. **(b)** Artificial stereo view after the random dots insertion, this panel represents the stereo view as it is given to the algorithm.

Artificially Created Right View

The steps to create the other view (right view) of each stereo images couple are the same of those explained above. Obviously the Gaussians positioning requires the knowledge of the disparities we want to give to the algorithm as inputs. Each blob position is chosen starting from the position of the corresponding blob in the left view shifted by the disparity value (horizontal shift) chosen for that specific blob. The global grid dimensions have, of course, to be left unchanged. Each blob between the two views can be left in its position (null disparity), shifted to the left (negative disparity) or shift to the right (positive disparity).

We also performed an experiment with a stereo pair of images containing only 3 blobs. This choice was made to test the algorithm on a scene in which each object had a disparity different from that of the others: in this case reducing from 7 to 3 the number of blobs allows to reduce the computational cost without loss of results validity.

4.1.2 Disparity Channel Results

In order to evaluate if the disparity channel algorithm is able to find depth discontinuity we performed a series of tests. To speed up the overall time of each test, we chose to never give as input to the algorithm more than three different disparities. When the disparities given as inputs are three we chose a set containing a positive value, a null value and a negative one.

We conducted a series of experiments to test if:

1. the algorithm is able to find the most salient blob when it is represented by a blob left alone at certain disparity;

2. algorithm results do not depend on the position of the salient blob;

3. algorithm results do not depend on the salient blob disparity value;

4. algorithm results do not depend on the number of salient blobs;

5. algorithm results are valid even when no blobs should be considered salient (no disparity value is specific to a single blob).

Table 4.2 presents an overview of the experiments in figure 4.3 and in figure 4.5. For each experiment are indicated: the disparity set, the disparities distribution (how many blobs for each disparity values), the number of the peaks in the output saliency map (SM) and the maximum value in the saliency map.

Test	Disparity Value	n° blobs	n° peaks	Maximum value
$1° - 7°$	-5	3	1	$[10000 - 12000]$
	0	1		
	$+5$	3		
$8°$	-5	6	1	17061
	0	0		
	$+5$	1		
$9°$	-5	1	1	10753
	0	3		
	$+5$	3		
$10°$	-5	1	2	12074
	0	5		
	$+5$	1		
$11°$	-5	2	no peak	3078
	0	2		
	$+5$	3		
$12°$	-5	5	no peak	3608
	0	0		
	$+5$	2		
$13°$	-5	1	3	12298
	0	1		
	$+5$	1		

Table 4.2: Blob disparities distribution tested.

Figure 3.16, at the end of the previous chapter, shows that the algorithm is able to find blob at an anomalous disparity.

Results do not depend on the position of the salient blob

In all the panels in figure 4.3 there are 3 blob of disparity -5, one of disparity 0 and 3 blobs of disparity +5. Hence, the 0 blob disparity is the most salient for all the 7 experiments shown in figure 4.3. Among the various tests conducted the only variable is the most salient blob (disparity zero blob) position. Figure 4.3 shows the ability of the algorithm to find object at an anomalous disparity regardless its position in the image. Peak values of saliency are comparable: they never drop below 10000 and never rise above 13000. The values at the non-salient blobs are also comparable to each other and are about 4000. Figure 4.4 is a three-dimensional representation of the saliency map resulting from the experiment with the salient

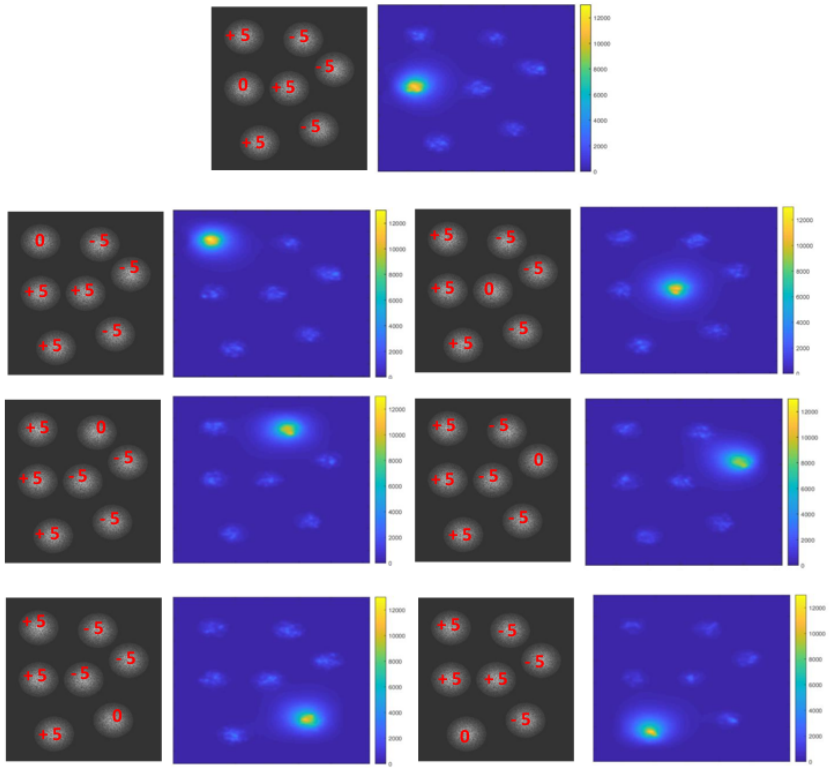
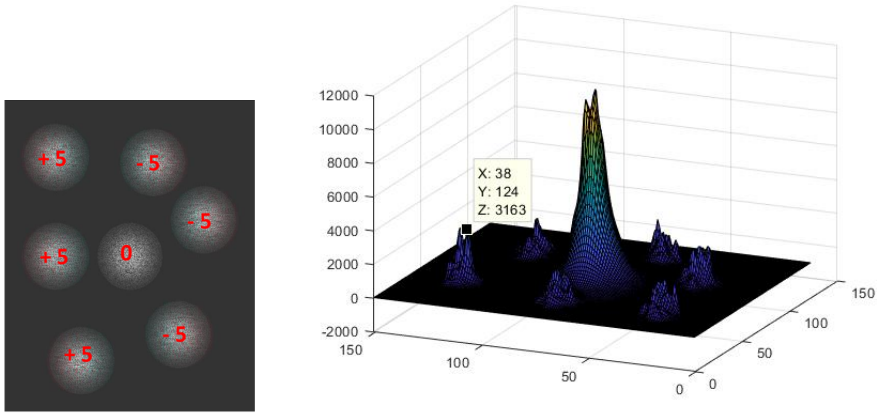


Figure 4.3: Disparity channel algorithm finds the most salient blob and its results do not depend on the position of the most salient blob. Each row shows single test input and output. Left panels are the left views given as inputs: red numbers indicate pixels of the horizontal shift. Right panels are the respective saliency maps returned as outputs by the disparity channel algorithm: the most salient areas are the yellow ones.

blob in the middle of the scene (second case shown in figure 4.3): the non-salient blobs present values that at maximum reach a quarter of the peak value (around 3000). The non-salient blobs are not supposed to have null values in the saliency map. The presence of an item on a homogeneous background causes the algorithm to recognize it and keep track of it in the output map, albeit with much lower values than the peak ones.

After had verified that the algorithm regardless of the position of the most salient blob is able to find it, we verified that algorithm performances were not influenced by the disparity value. We performed new experiments without changing the disparity set but only the disparity value of the most salient blob.



(a) *Input Images Red-cyan anaglyph*

(b) *Disparity Saliency Map: 3D representation*

Figure 4.4: Saliency map from Disparity Channel algorithm. **(a)** Red-cyan anaglyph from the two stereo images given as input to the disparity channel: red numbers indicate the horizontal shift relative to each blob and expressed as pixels number. The only blob at null disparity (central blob) is supposed to be most salient one. **(b)** Three dimensional saliency map representation. In the saliency map the most salient blob present a value around 12000 while never have values above 3500.

Results do not depend on the salient blob disparity value

As shown in figure 4.5 the algorithm can find a salient blob even when its disparity is either +5 (first row in figure 4.5) or -5 (second row in figure 4.5). In the first case the disparities set is reduced because there is no blobs at null disparity and the saliency map shows a peak value for the most salient blob that is higher than the ones shown when the disparity set is bigger. Even if we do not have enough information about human biology behavior, this result could be reflected in the mechanism of selective attention. If the number of depth planes (figure 4.1) on which the objects are arranged on the scene is reduced, the visual search is dispersed over a small space so the object at an anomalous disparity will attract attention in a shorter time, resulting even more salient than the surrounding objects. In the saliency maps of both cases the non-salient blobs values are around one quart of the peak values as the case show in figure 4.4. The results remain unchanged whether the algorithm is required to look for objects with zero disparity, or it is not required to (zero disparity is not included in the input set).

Results do not depend on the number of salient blobs

When the blobs left at an anomalous disparity are more than one we have to distinguish two cases:

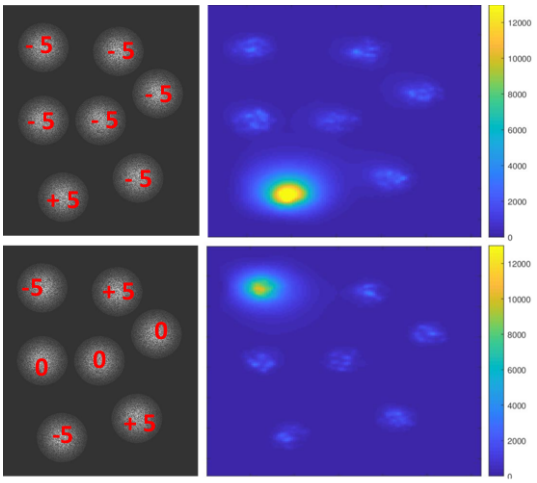
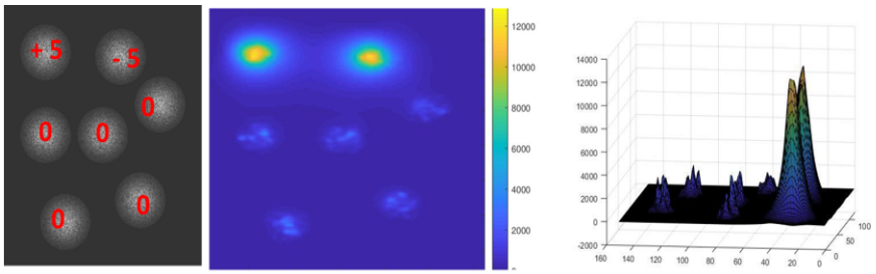


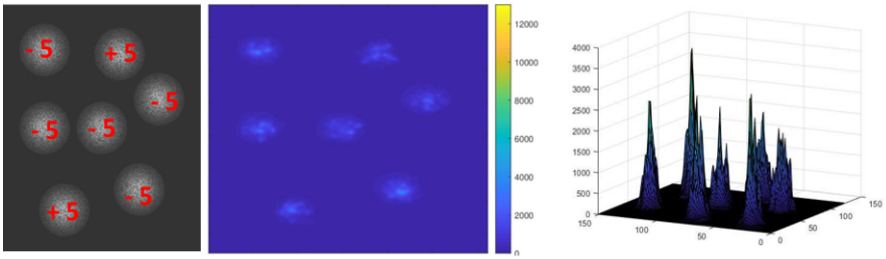
Figure 4.5: Disparity channel algorithm finds the most salient blob and its results do not depend on the disparity value of the most salient blob. Each row shows single test input and output. Left panels are the left views given as inputs: red numbers indicate pixels of the horizontal shift. Right panels are the respective saliency maps returned as outputs by the disparity channel algorithm: the most salient areas are the yellow ones.

- two or more blobs are characterized by an anomalous disparity and these disparities are different from each other (figure 4.6(a));
- two or more blobs are characterized by the same anomalous disparity (figure 4.6(b)).

In the first case shown in figure 4.6(a) the input stereo images pair presents one blob at disparity -5, 5 blobs at null disparity and one blob at disparity +5. Both this two blobs have the same level of saliency and in fact in the respective output map there are two peaks of similar values (around 12000) and 5 areas whose value on the map is always around 3500. The experiment shown in figure 4.6(b) represents the output in the case where in the pair of stereo input images there are 7 blobs distributed in an in-homogeneous manner between two disparities (5 blobs at disparity -5 and 2 blobs at disparity +5). Although it is ecpected to find two higher peaks corresponding to the displaced +5 blobs, te output map does not indicate the presence of actual peak values: the highest values are around 4000 and are found for each blob present in the scene. The comparison between the two cases in figure 4.6 suggests that the disparity algorithm is able to find blobs at anomalous disparities when there is a single blob for that specific disparity: this is confirmed by the results shown in figure 4.7 where for each disparity in the scene there is only a blob. If each blob in the scene has a unique disparity value the algorithm recognized each one of them as salient. In figure 4.7, even if with values slightly different, each blob area corresponds to a peak in the output saliency map (SM). The fact that the algorithm



(a) *2 blobs for 2 anomalous disparities*



(b) *2 blobs for one anomalous disparity*

Figure 4.6: Results and salient blobs number.**(a)** from left to right: input images (red numbers: horizontal shifts for each blob), two-dimensional saliency map: max value around 12074, most active areas are the yellow ones, three-dimensional saliency map: non-salient blobs have all values in between 2000 and 3500. **(b)**from left to right: input images (red numbers: horizontal shifts for each blob), two-dimensional saliency map: max value around 3608, most active areas are the yellow ones, three-dimensional saliency map: non-salient blobs have all values in between 2000 and 3500

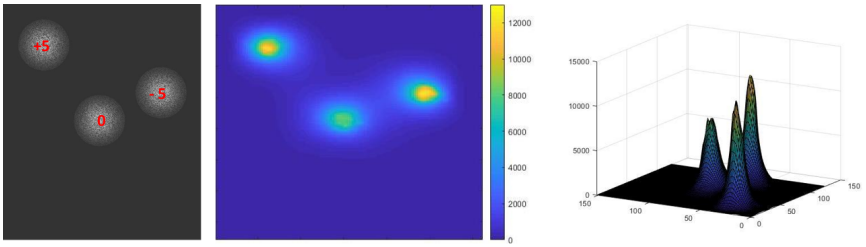


Figure 4.7: Each blob has a unique disparity. From left to right:Input left view (red numbers: horizontal shifts for each blob), two-dimensional saliency map: max value around 12298, most active areas are the yellow ones, three-dimensional saliency map:blobs have all values in between 8000 and 13000.

do not recognize as salient two items at the same anomalous disparity is probably due to the normalization part of the algorithm. It is undoubtedly a question that must be studied and resolved. To do this it is also necessary to collect a series of visual search experiments results, in order to better understand how human attention is distributed in these cases. Meanwhile, seeing the good results obtained in the other

cases studied, we decided not to change the algorithm taking into account that the disparity channel real contribution to the processing of the three-dimensional global salience map occurs when objects lie alone on an anomalous depth plane (figure 4.1).

Results when no anomalous disparity are present

Figure 4.8 shows the last situation we tested: a scene where blobs are equally distributed among all the disparities in the input set. Two blobs are at disparity -5, two blobs are at null disparity and three blobs are at disparity +5. In this case the algorithm is not rightly able to find a real salient blob: maximum values is around 3600 comparable with al the values find at non-salient blobs in all the previous cases.

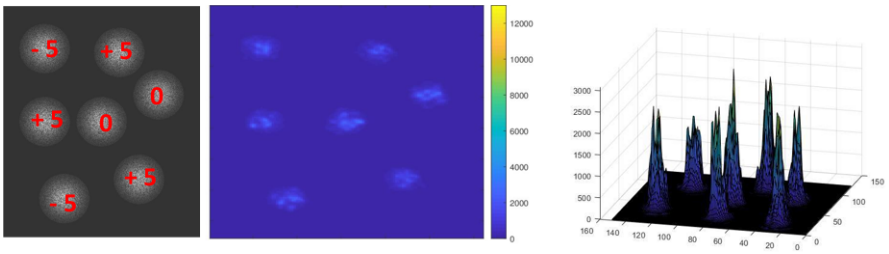


Figure 4.8: Disparity channel results when no anomalous disparity. From left to right: Input left view (red numbers: horizontal shifts for each blob), Two-dimensional saliency map: max value around 3600, most active areas are the yellow ones. Three-dimensional saliency map: blobs have values below 3500

At the end of these first tests, based on the sufficiently good results, we decided to continue testing the whole model using real scene as inputs, in order to see if introducing notion about stereo disparities distribution model performances change and hopefully improve. To evaluate the final saliency map a dataset with both stereo views and a valid ground truth (i.e. fixation density map estimated using an eye-tracker) are needed.

4.2 Real Scene Images: Three-dimensional model Results

To evaluate the performance of the extended proto-object based saliency algorithm, the ability of the model that includes disparity discontinuities notion to predict human fixations points is compared to that of the model without depth information. Any differences between the three-dimensional output saliency map from the new

version of the computational model, and the two-dimensional output saliency map from the old version of the computational model is purely a result of adding information about three-dimensional because the only difference between the two is the presence of the new Disparity Channel. For all the following tests we used the image database GAZE3D [16] that is right now the only online accessible database complete with all the information we need to be provided as inputs and to be used as ground truth. We anticipate that the small number of picture in the dataset will be a limitation for the quantitative evaluation of the model, explained in chapter 5.

4.2.1 Dataset Description

The GAZE3D dataset contains 18 stereoscopic images and the associated fixation density map, disparity map, depth map, and the raw eye tracking data.

Three-dimensional Images are provided as a stereoscopic couple, the two stereoscopic views are used as inputs to test our model.

Disparity Maps are used to extract the disparities set to give as input to the disparity channel algorithm (figure 3.9).

Fixation Density Maps are used as ground truth to evaluate the model through distribution based evaluation metrics (chapter 5).

Raw eye tracking data used to extract binary fixation maps to use as inputs with the location based evaluation metrics (chapter 5).

Three-dimensional Images

The dataset includes 18 3D images composed by 2 2D PNG images. The 2 PNG images that compose a stereoscopic couple are marked as left ("_L") and right ("_R") image and they are used as inputs for the model. A briefly description fo the sources is shown in table 4.3. The database is provided free from charge and it is composed by 18 stereo pairs: the first 10 come from the Middlebury 2005/2006 database [29], and the other 8 images are acquired at the University of Nantes campus [36]. All the stereo images are rectified (figure 3.10) but no information about the rectification method are provided. Reference for the original sources is the whole first column of figure 4.13.

Source	Resolution	Description	Author
1	1278×1080	A painting artist desk	Middlebury College
2	1278×1080	Some dolls on a desk	Middlebury College
3	1282×1080	Some books on a desk	Middlebury College
4	1279×1080	A baby on in front of a map	Middlebury College
5	1228×1080	Some cleaning stuff on a desk	Middlebury College
6	1274×1080	Middlebury merchandising	Middlebury College
7	1286×1080	Colored and shaped objects	Middlebury College
8	1194×1080	Some plastic objects	Middlebury College
9	1247×1080	Objects at different planes	Middlebury College
10	1192×1080	Some rocks	Middlebury College
11	1920×1080	Two boxers training	IRCCyN
12	1920×1080	Hall between two buildings	IRCCyN
13	1920×1080	Girls in a chemical laboratory	IRCCyN
14	1920×1080	Two men reporting	IRCCyN
15	1920×1080	A man speaking on phone	IRCCyN
16	1920×1080	Two men playing football	IRCCyN
17	1920×1080	Tree branches slow moving	IRCCyN
18	1920×1080	A man using an umbrella	IRCCyN

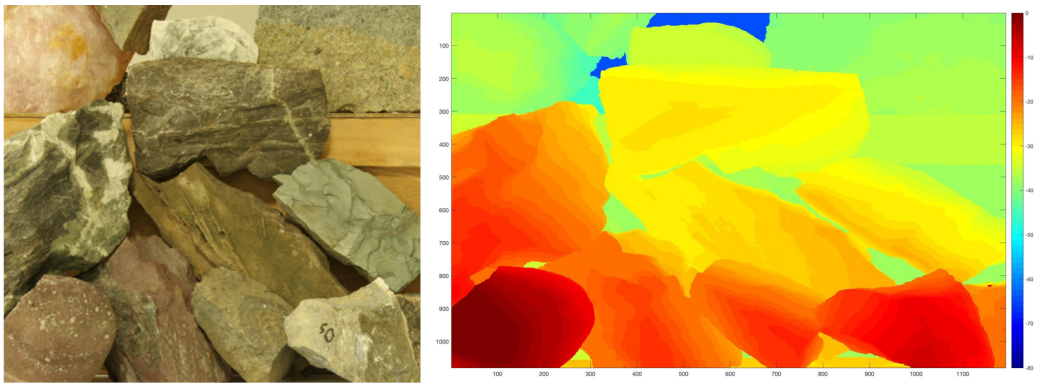
Table 4.3: GAZE3D dataset sources description.

Disparity Maps

In the dataset for each stereoscopic pair there is its associated Disparity Map store like an image in a matrix in a size of:

$$\text{height of the image} * \text{width of the image}$$

To each pixel is assigned a value that indicates the disparity existing between the left and the right projections of the point that in the visual scene, corresponds to the pixel in question. Due to the fact that the stereo pairs are composed by two rectified views (see figure 3.10 and section 3.2.1), such disparities indicate the number of horizontal shift between the left and the right representation. From the disparity maps, information about the whole disparity set of a specific stereo couple can be extracted, such values can be use as input for the Disparity Channel in the model (figure 3.9). Figure 4.9 has been created using MATLAB2017b to show an example of one disparity map from the dataset that was stored in a matrix.



(a) *Image n° 10: left view* (b) *Image n°10: Disparity Map*

Figure 4.9: Image n°10 from GAZE3D dataset. **(a)** Original left view, size: 1192x1080. **(b)** Disparity Map associated to Image n°10: the column on the right gives and indication of the disparity values.

Fixation Density Maps and Binary Fixation Maps

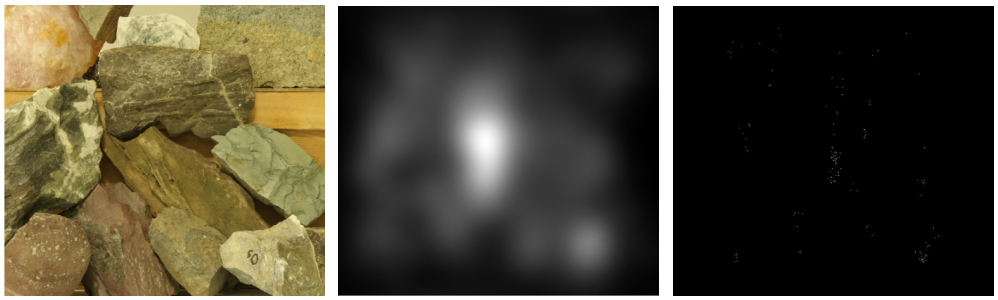
As well as the disparity map for each of the 18 stereo pairs the dataset provides the Fixation Density Map (FDM). These maps are saved as PNG black and white images where more a pixel is white, more this pixel is visualized by the observers. The fixation density maps are obtained through eye tracking experiment which are based on the link between visual attention and eye movement. An eye-tracker exploits the natural reflection of infrared light on the human eye. Each eye-tracker is equipped with an infrared emitter directed towards the eye (or eyes) of the observer and a sensor that detects the reflected ray, thus being able to extract the fixation point on a screen. A fixation point is defined as a point that occupies a location where the gaze stopped from 2 to 4 tenths of a second. In the case of these dataset a series of participants is required to look at a screen while a three-dimensional scene from the dataset is reproduced, the eye-tracker is able to track the eyes movement, the result pattern is then processed to obtain the FDM of the single observer and once the experiment has been completed on all the participants, the average FDM is calculated to obtain the final fixation density map of which an example is shown in figure 4.10(b). The whole set of Fixation Density Maps is shown in the second column of figure 4.13.

The real output of the eye-trackers is a collection of positions within the image under analysis (the one that the observers are asked to look at) that indicates the coordinates of fixated points. When the Fixation Density Maps are given as a uniform distribution of white areas on a black background il means that they have been smoothed through some post processing operations. The most common post processing operation includes the convolution with a two-dimensional gaussian that generally has a standard deviation (in term of pixel number) equal to the the

standard deviation ot the error of the eye-tracker. This operation acts as a regulation in order to include in the FDM information [4]:

- **Eye-tracker error**;
- **Uncertainty about what the observer is really looking at**.

This smoothing can affect the evaluation metrics results (chapther 5). In the case of the GAZE3D the FDMs are provided already smoothed but no information about the eye-tracker error and its standard deviation is provided. Luckily the raw eye-tracker data include the fixated points position for all the images in the dataset so it is possible to rebuild a binary fixation density map as a black map (null pixel value) where only the pixels corresponding to fixated locations are forced to be white (pixel value equal to 1). Example is shown in figure 4.10(c).



(a) *Image n° 10: left view* (b) *Image n°10: FDM* (c) *Image n°10: binary FM*

Figure 4.10: Image n°10 from GAZE3D dataset. **(a)** Original left view. **(b)** Fixation density map linked to image n°10: more white pixels are the more visualized by the observers **(c)** Binary fixation map: white pixels indicate the fixated points positions relative to image 10. These positions are stored as eye-tracker row data and shown using MATLAB2017b.

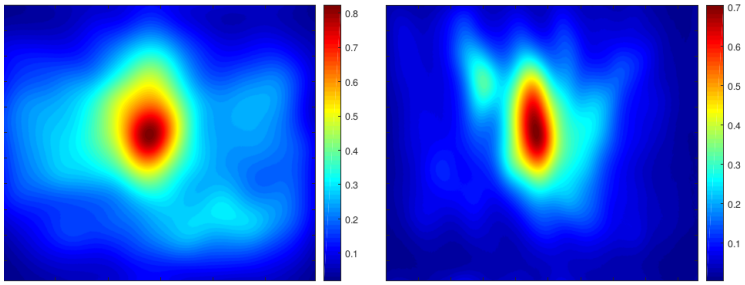
The eye-tracker works directly in binocular mode and the image shown to the participants is the three-dimensional combination of the two stereoscopic view with the task to "look around the image as you naturally would". The eye-tracker is the SMI Hi-Speed, information about the test and the observers are provided in table 4.4.

Observer distance	Duration	Pre-screening	N° observers	Age:[Range]
63cm	10 seconds	Snellen, Ishihara	35	[18 46]

Table 4.4: Eye-tracking tests conditions.

4.2.2 Center Bias Problem

As visible in the example in figure 4.10 and in the second column of figure 4.13 human fixation points have a natural center bias. It means than in front of any scene an observer tends, especially for the first few seconds of testing to fix the center of the screen which is the optimal point of view. In the same way, even the photographer who takes the picture is inclined to center the subject in the center of the image. This phenomenon can not be predicted by any of the computational models for visual salience. This phenomenon is visible if one observes the figure 4.11 which shows the average of the fixation density maps of the images coming from the Mildelbury dataset (a), first 10 images in table 4.3, and those acquired in the University of Nantes campus (b), last 8 images in table 4.3.



(a) *Average of FDM: from Image n°1 to Image n° 10*

(b) *Average of FDM: from Image n°11 to Image n° 18*

Figure 4.11: Fixation density Map center bias. Averaging the the saliency maps over all the images respectively belonging ti one of the two groups, is clearly visible that the more observed area appears to be always in the center of the scene. **(a)** average of the first 10 FDMs dataset. **(b)** average of the last 8 FDMs of the dataset.

Some evaluation metrics (chapter 5) we used to measure the ability of our computational stereo vision model of proto-object based saliency to predict the location of fixated points, are strongly influenced by the occurrence of the center bias phenomenon. For that reason we decided to reduce its influence by subtracting to each group of fixation density map (1st group: from image n°1 to image n°10, second group: from image n°11 to image n°18) the averaged obtain across all the fixation density maps of the corresponding group. The map subtracted to the first group is shown in figure 4.11(a) and the one subtracted from the second group is shown in figure 4.11(b). Results are shown in figure 4.13 in the third column.

Concerning the binary fixation density maps (figure 4.10(c)) the center bias is removed forcing to 0 all the values include in the yellow and red areas in figure 4.11. Thus for both groups of fixation binary maps we will not consider as fixated points all the points that correspond to a pixel whose value, in the respective map obtained through the average of all the fixation maps of the corresponding group, is included

in the range $[0.5 - 1]$ (figure 4.11(a)(b) for reference).

4.2.3 Proto-Object Based Saliency Map in Three-dimensional Space

In order to test the new extended proto-object based visual saliency model it is necessary to exactly know which are the inputs to give to the model and to the new disparity channel. Input images are time by time the two stereoscopic views corresponding to one of the 18 scenes of the GAZE3D dataset (shown in the first column in figure 4.13 and described in table 4.3).

Choice of Input Images

The intensity, color opponency and orientation channels use the view stored as "right image" for evaluating the two-dimensional saliency map while the disparity channel required both the left and the right view. To chose the reference view for the evaluation of the SAD maps (section 3.2.4 we observed the disparity values set: all of them are composed by number of negative disparities that is much higher than the number of the positive ones. This, combined with the fact that the images are all rectified, directed us to choose the one indicated as "right image" as the reference one to start the algorithm for the SAD maps creation. As confirmation of the correctness of the choice made, for one of the 18 scenes described in the table 4.3, we build the SAD maps for the most common disparity value in the whole set (example in figure 4.12), choosing once the left view and once the right view as a reference. The SAD map evaluated using the left view as reference did not show any pixel at the analyzed disparity confirming that the right choice is to set the right views as the reference for the SAD map evaluation. We noticed that Image n°12 has only positive disparity so we repeat the steps for this image. Also the study on this image confirmed what has already been discovered.

Choice of Disparity set

The disparity channel algorithm required as input a set of plausible disparities to look for between the two stereo views of the same scene. As visible from table 4.5 disparity set dimension varies from one image to another and it can reach up to more than 100 different disparities for a single image from the dataset. From equation 3.4 we know that for a total of N disparity values the opponency maps to give as inputs to the grouping algorithm are $(N) * (N - 1)$, this means that the time to complete a test increases greatly as the number of disparities taken into consideration increases. In order to limit the computational cost for each stereo pair image we observe:

- **Maximum disparity value**;

- **Minimum disparity value**;
- **Most common disparity value**.

This three disparity values are shown in table 4.5. After that, if the most common value is close to the central value, we choose nine value of disparity: four below the most common value, four above the most common value and the most common value itself (figure 4.12). Anyway as visible from the summary in table 4.5 we never considered more than ten different disparity values. This does not let us to see items at each disparity but due to the fact that an object extends over several levels of neighboring disparity (see figure 4.9), the hypothesis on which the algorithm is based does not lose validity. Hopefully at least part of each object is visible and will contribute to the grouping mechanism.

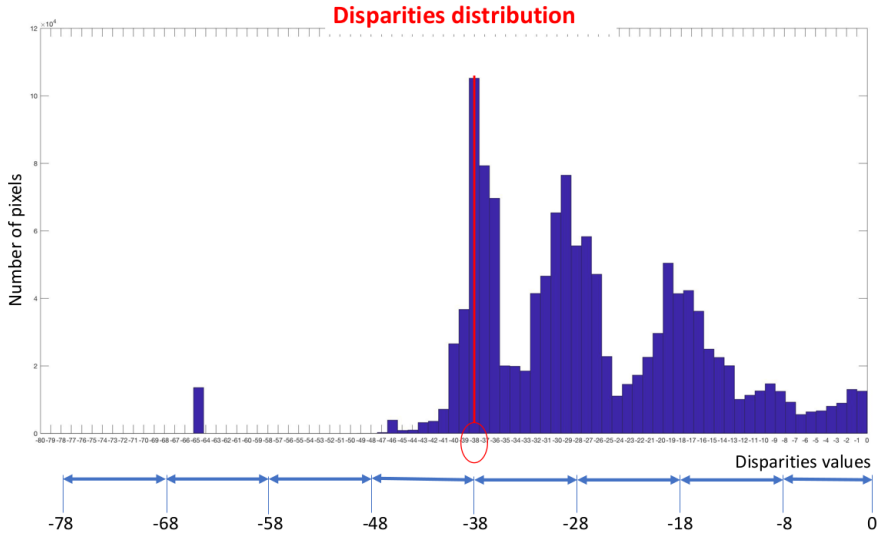


Figure 4.12: Disparity distribution for the scene n°10. Value marked with the red circle indicates the most common disparity in the scene. Values reported under the histogram are those chosen to test the model when using image n°10 as input.

Figure 4.13 shows the saliency maps obtained from the 18 tests performed on the images of the considered dataset and the relative fixation density map with and without center bias. To make the changes introduced by the addition of the disparity channel when we summed up the two-dimensional saliency map and the saliency map obtained from the disparity channel we used a 1:4 ration. It means that the weight given to the map obtained through the disparity channel is equal to 80%. This choice is in contrast to what was stated in chapter 3 where we affirmed that two-dimensional features have more relevance in determining the salience map, considering a more biologically plausible model, but this is made only in order to better show changes introduced when depth discontinuities information are included

Source	Disparities Range	Most common	Range for saliency
1	[-112 - -3]	-72	[-112:10:-72;-55:17:-4]
2	[-23 - 20]	-20	[-23;-20;-10;0;10;20]
3	[-75 - 1]	-69	[-75;-69;-54;-39;-24;-9;1]
4	[-80 - 0]	-34	[-80;-70:12:-10;2]
5	[-96 - 0]	-59	[-96;-89;-74;-59;-44;-29;-14;0]
6	[-86 - 0]	-73	[-86;-73;-58;-43;-28;-13;0]
7	[-80 - 0]	-54	[-80;-69;-54;-39;-24;-9;0]
8	[-63 - 22]	-10	[-63;-50;-40;-30;-20;-10;0;10;22]
9	[-63 - 5]	-61	[-63;-61;-46;-31;-16;5]
10	[-80 - 0]	-38	[-80;-68;-58;-48;-38;-28;-18;-8;0]
11	[-7 - 14]	6	[-7;-2;2;6;10;14]
12	[1 - 17]	4	[1;4;7;10;13;17]
13	[-44 - 26]	-41	[-44;-41;-26;-11;4;19;26]
14	[-6 - 39]	-5	[-5;5;15;25;39]
15	[-18 - 32]	17	[-18;-13;-3;7;17;32]
16	[-20 - 22]	11	[-20;-9;1;11;22]
17	[-24 - 19]	-3	[-24;-13;-3;7;19]
18	[-52 - 15]	-38	[-52;-38;-23;-8;7;15]

Table 4.5: Disparity sets description.

in the model. From a first visual comparison it is noted that none of the saliency map (the two-dimensional map, the one processed by the disparity channel and the three-dimensional map) is able to predict the presence of the center bias that characterizes the fixation density maps as provided by the dataset. Confronting the output saliency map a change introduced by the proposed extension of the two dimensional model is clearly visible. Thus we can already affirm that information about depth discontinuity change the prediction of fixated points. To know if this changes improve the prediction quality it is necessary to evaluate the two model (the two-dimensional and the three-dimensional) using some of the most common evaluation metrics for models of visual saliency. For that reason, next chapter is entirely dedicated to the description of these metrics and the discussion of the results obtained when they are utilized to quantify the goodness of the salient areas prediction operated by the two-dimensional and the three-dimensional proto-object saliency models.

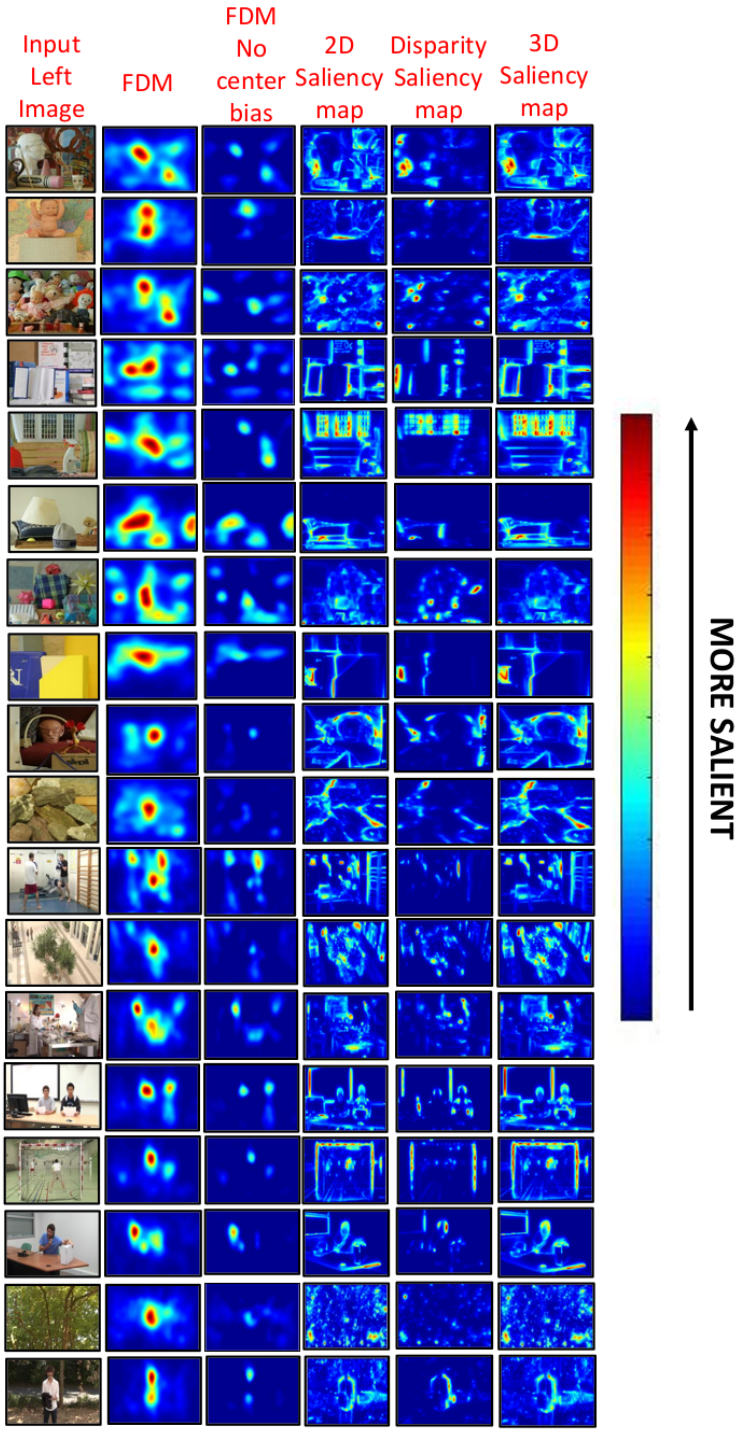


Figure 4.13: All the 18 left views from GAZE3D dataset stereo pairs, their associated eye-fixation maps, fixation density maps after center bias elimination, saliency maps calculated using the previous two-dimensional model, saliency maps calculated using the disparity channel algorithm and the saliency maps calculated using the new computational stereo-vision model of proto-object based saliency in three-dimensional space.

Chapter 5

Model Evaluation

A simple visual comparison between the various salience maps and the fixation density map is not sufficient to estimate how the approximation of human visual fixations is correct and valid. For that reason evaluation metrics are need to calculate an index of the degree of similarity (or dissimilarity) between any ground truth and a salience map elaborated by a specific computational model. Finding the correct method to evaluate the quality of a predicted saliency map, has not been easy since when the first model of visual attention came out. The choice changes in respect of a lot of parameters and the overall evaluation setup (eventual tasks for the visual search experiment, number of participants, distance from the screen, etc.) plays a decisive role. As all the evaluation metrics the ones used for saliency, as affirmed by Wilming et al., should respond to some basic needs [45]:

- **Few parameter**;
- **Intuitive scale**;
- **Low data demand**;
- **Robustness**.

The metrics reported in this chapter are all already widely used to evaluate the quality of models of visual attention, there are many publications that evaluate their performance so we can say that they meet these basic requirements.

In general, when possible, it is better to evaluate the model with more than one type of evaluation metrics: the first evaluation shows a series of results, the subsequent confirm, or deny, that the trend obtained is correct. Moreover, repeating the same measure with different metrics allows to correct the eventually first choice of metric unsuitable for the evaluation setup in question. Knowing that the fixation density maps of the dataset we used were obtained through experiments in which no specific task was given to the participants, makes our case less restrictive the regarding the evaluation metrics choice. Generally a determining factor is the type

of fixation map that are available but, again, in our case the dataset give us both the map of fixation locations (example in figure 4.10(c)) and the continuous fixation map (examples in figure 4.10(b)) allowing the use of a broad set of evaluation metrics. In fact metrics are divided in: location based and distribution based. The former require a discrete input in the form of fixation locations map while the letter require a continue distribution as the one of our blurred fixation density maps.

We have chosen three of the most used metrics that are also the same ones used by Russel et al. to evaluate the model's performance based on the concept of proto-object of which we are expanding: **Area Under the ROC Curve** (AUC) and **Kullback-Leibler divergence** (KLD) and **Normalized Scanpath Saliency** (NSS). Two of these metrics have been adapted to saliency maps evaluation from other field of study: AUC is adapted from signal detection and KLD is used in information theory. NSS has been designed specially for saliency maps. Table 5.1 summarizes the main differences between the three. Referring to table 5.1, a similarity metric is a metric that measures the degree od similarity between a ground truth map and a prediction so a high score from these metrics means that the two maps are similar while for the dissimilarity metrics the situation is the opposite (measure of dissimilarity: high scores mean the two maps look different).

Metrics	Similarity	Dissimilarity
Location Based	AUC, NSS	
Distribution Based		KLD

Table 5.1: Main characteristics of the chosen metrics.

The code used for all the metrics is available online on the MIT Saliency Benchmark webpage [17] in the form of MATLAB code.

5.1 Receiver Operating Characteristic (ROC) curve

In decision theory the ROC (Receiver Operating Characteristic) curves are used to study the relationship between true and false positive. Calculating the area under the ROC curve (Area Under Curve, AUC) is possible to quantify the ability of the visual saliency model under analysis to predict human fixation points as they appear in the ground truth. The AUC is the most diffused method to evaluate visual saliency models performance. In order to build a ROC curve, we need to interpret the prediction of possible fixation points as a binary classification. This can be done setting a discrimination threshold and analysis the saliency map values, corresponding to the fixation points, relatively to this threshold. The information on the fixation points position is given by the fixation map with discrete values

where all this locations are marked with a white pixel. This confirms the statement that AUC is a location based metric. Depending on the way that false positive are counted, different methods for calculating AUC are distinguished. In our case to define if a point that is marked as salient by the model, is a true positive (TP) or a false positive (FP) the following steps must be followed.

1. **A threshold value is settled**;
2. **SM (Saliency Map) points corresponding to the fixation points are selected**;
3. **A set of random (Saliency Map) SM points is selected**: excluding al the pixels already selected (each pixel has to be selected once);
4. **Each one of the selected points is marked as TP or FP**:
   considering a point p in the saliency map at the position (x_i, y_i)
   $$\text{p} = \begin{cases} \text{VP}; & \text{if SM}(x_i, y_i) > threshold \wedge p \text{ is a fixation point} \\ \text{FP}; & \text{if SM}(x_i, y_i) > threshold \wedge p \text{ is NOT a fixation point} \end{cases}$$
5. **Threshold is changed**: threshold range goes from the minimum value of the SM up to the maximum value of the SM;
6. **Steps 2,3,4 are repeated whit a set of different thresholds**.

The number of TP and FP is stored for each threshold value so at the end an ROC curve can be built: along the two axes we can represent sensitivity and (1-specificity), as True Positive Rate (TPR, fraction of true positives)

$$\text{TPR} = \frac{TP}{\text{total n° fixations}} \tag{5.1}$$

and False Positive Rate (FPR, fraction of false positives)

$$\text{FPR} = \frac{FP}{\text{FP} + \text{total n° not fixations}} \tag{5.2}$$

For each threshold value a point of the curve is defined. The AUC values is in between 0 and 1:

- $0 \leq AUC < 0.5 \rightarrow$ algorithm perfoms worse than chance;
- $AUC = 0.5 \rightarrow$ algorithm predicts eye fixations at chance;
- $0.5 < AUC < 1 \rightarrow$ algorithm perfoms better than chance;

- $AUC = 1 \rightarrow$ algorithm perfectly predicts eye fixations;

Looking at the definition of both TPR and FPR from equations 5.1 and 5.2 it is clear that the presence of center bias (see chapter 4.2.1 and figure 4.11) in the ground truth data of the eye-tracker. The natural human tendency to look at the center of a screen when an image is shown as well as the tendency of the photographer to put the object in the center of the picture can not be predicted by any of the models proposed by saliency models state of the art. For that reason we decided to use the AUC metric also to compare the salience maps obtained as outputs from the two-dimensional and three-dimensional models and fixation maps (in this case we refer to the binary maps before the shading) after the removal of the center bias (as proposed in the section 4.2.1). Scores are shown in tables 5.2 and tables 5.3

5.2 Kullback–Leibler Divergence (KLD)

In probability theory and in information theory, Kullback–Leibler divergence (KLD) is a non-symmetric measure of the difference between two probability distributions P and Q. If denoted by $\mathrm{D}_{KL}(P||Q)$, it is the measure of information lost when Q is used to approximate P. The Kullback–Leibler Divergence in the discrete case is defined as:

$$\mathrm{D}_{KL}(P||Q) = \sum_i \mathrm{P}_i log\left(\epsilon + \frac{\mathrm{P}_i}{\epsilon + \mathrm{Q}_i}\right) \tag{5.3}$$

where ϵ is a regularization constant. Thus considering Q our saliency maps and P the fixation density map we want to approximate, a good approximation quality is indicated by small KLD cores, for that reason as shown in table 5.1 is included in the dissimilarity evaluation metrics group. Therefor we have:

- low KLD $\rightarrow$ saliency map performs better than chance;
- higt KLD $\rightarrow$ saliency map performs at chance.

As KLD metric is distribution-based, it required as input the fixation density map represented as a probability distribution: after blurring the fixation locations (figure 4.10(b)) into a fixation density map (figure 4.10(c)). From equation 5.3 we can affirm that when in the fixation density map (P in the equation 5.3) there is a pixel whose value is not zero but the corresponding pixel in the saliency map (Q in the equation 5.3) has a null value or a value close to zero a large quantity is added to the global KLD score. This occurs in the areas affected by the center bias for that reason also for KLD evaluation metrics we tested both the similarity between the fixation density map as it is and the saliency map and the similarity between the fixation density map after the center bias removal and the saliency map. Results for our models are shown in tables 5.2 and 5.3.

5.3 Normalized Scanpath Saliency (NSS)

The Normalized Scanpath Saliency (NSS) was specifically introduced in 2005 to evaluate human saliency model ability to predict fixation locations. It computes the average normalized saliency value at fixation locations. Given a saliency map Q and the fixation binary map P the NSS is evaluated as follow:

$$\mathrm{NSS}(P||Q) = \frac{1}{N}\sum_i \bar{Q}_i \times P_i \tag{5.4}$$

where

$$N = \sum_i \mathrm{P}_i \text{ and } \bar{Q} = \frac{Q - \mu(Q)}{\sigma(Q)}$$

. So $\bar{Q}$ is the normalized saliency map that have a zero mean and a unit standard deviation, N is the total number of fixation points. NSS score can be null, positive or negative.

- $NSS = 0 \rightarrow$ saliency map performs at chance;
- $NSS > 0 \rightarrow$ correlation between salient points and eye fixations is greater than chance;
- $NSS = 0 \rightarrow$ anti-correspondence between saliency map and eye fixation data;

NSS evaluation metric is used, as the AUC and KLD, to compare the performances of the computational model of proto-object based saliency that includes the extension with the depth information, with the performances of the pure two-dimensional model. If adding the disparity channel, the evaluation metrics scores denote that model move away from the performance at chance conditions, it means that extension has improved the model's ability to predict the location of fixation points.

5.4 Improvement against Two-dimensional Proto-Object based saliency

The model is evaluated by comparing the saliency map generated by our model with the ground truth data in the form of human fixations. Evaluation metrics scores are calculated for each one of the 18 images in the GAZE3D [16] [29] [36]. Results presented in both table 5.2 and table 5.3 are obtained by averaging the single results, one for each image in the dataset, over all the sources available. We decided to repeat the same evaluations changing the inputs. In table 5.2 results are obtained comparing saliency maps with the fixation density map (or fixation binary map see figure 4.10(b) and figure 4.10(c)) as it is given by the dataset while for the results

in table 5.3 we used the fixation maps without the center bias and we convoluted with a two-dimensional Gaussian with a standard deviation of 6 pixels the saliency maps. This last step was carried out in order to take into account the standard deviation of the eye-tracker error. This error standard deviation is known once the type of eye-tracker is known (information provided by those who have spread the dataset online). It is actually provided in terms of degrees of visual field but the equivalent pixel value can be estimated knowing the distance from the screen, the size and resolution of the screen itself. Obviously the value changes with the size of the image shown but it is assumed, without loss of generality, that the images are all 1920×1080 and shown in full screen. Below the list of the information obtained from the data provided by the dataset, whose download includes a pdf attachment containing information on the instrumentation used.

- **Eye-tracker:** SMI Hi-Speed → **error standard deviation:** 0.25°;
- **Monitor:** Panasonic BT-3DL2550S → **screen dimension:** 25.5″, **screen resolution:** Full HD (1920x1080 pixels).

Starting from these data we evaluate an approximate value of the eye-tracker error standard deviation that is equal to 6 pixels. So table 5.3 shows models performances obtained after correcting some errors introduced by the eye-tracker: center bias and inaccuracy in the definition of the coordinates of the fixation point (whose value is included in the user manual, available online).

Looking at the results from all the metrics and for both test condition it is immediately visible that, even if it is small, the introduction of three-dimensional information, causes a global improvement on the model performances.

If eliminating the center bias and smoothing the saliency maps do not really change AUC scores the other two metrics scores undergo an important change as a result of these operations. These changes are, in our opinion, essentially related to the removal of the center bias. In fact the center bias effect adds fixation points that can not be found from neither two and three dimensional model and this causes an increase in the total KLD score (as explain in section 5.2 and deducible from equation 5.3) and a decrease in NSS. From the definition of the NSS (equation 5.3), shown in equation 5.4 it is easy to understand that, increasing the number of total fixation points (consequence of the effect of center bias), the total NSS score decreases.

The fact that the improvements introduced, following the extension of the algorithm, are contained, may be due to a number of factors, first of all the disproportion between the multiplicative weights that are used to linearly sum the two-dimensional saliency map and that obtained from the evaluation of the binocular disparities. In fact, as anticipated at the end of chapter 3, is more biologically plausible to consider that the overall two-dimensional features contribute more to the final saliency map. Moreover, as affirmed by Hu et al. [13] and previous works, when, for the creation of the ground truth, the participants are left to observe the image for a relatively

long time, the two-dimensional features acquire more importance. This happens for observation times longer than $5s$, as in the case of our dataset (table 4.4).

Our disparity channel makes no distinction based on the disparity range. As shown in picture 4.1 each disparity value is associated with a specific distance from the observer, big disparities mean small distances and objects that appear bigger to the observer. The objects characterized by big disparity values may be particularly salient as a result of the grouping mechanism. So somehow we would have to split the dataset into multiple subgroups based on the range of binocular disparities associated with the source to better notice the improvements introduced through the model extension. This would have been possible if the starting dataset had been wider. In fact, already with 18 images it was not possible to reach statistical significance.

Model	AUC	KLD	NSS
Pure 2D model	0.591	0.927	0.293
2D model + disparity	0.595	0.917	0.324

Table 5.2: Evaluation metrics results:no smoothed SM and FDM with center bias.

Model	AUC	KLD	NSS
Pure 2D model	0.633	0.672	0.512
2D model + disparity	0.634	0.667	0.516

Table 5.3: Evaluation metrics results: smoothed SM and FDM without center bias.

5.4.1 Statistical Significance

Comparing the scores obtained from the different evaluation metrics, we have ascertained that introducing notions on depth discontinuity improves the quality of the salient points prediction made by a computational model that exploits the notion of proto-object rather than the feature integration theory. To state that the results obtained are not the result of the chance, it is necessary to carry out a comparative test that will give back information about the statistical significance. Significant was calculated using a paired t-test between the results of proto-object based saliency map in two-dimensional space and the results of proto-object based saliency map in three-dimensional space. We performed a two-tails, paired Student t-tests, with a significance level of $\alpha = 0.05$. For all the tests, results never reach the significant at the chosen level. All the $p - values$ for both cases (with and without center bias and before and after saliency map smoothing) turned out to be around 0.2.

This does not allow us to deny the null hypothesis according to which each score difference between the two-dimensional and three-dimensional models is a random result. Previous works [13] have already shown that paired t-tests carried out on the results obtained from experiments conducted starting from GAZE3D dataset were not able to show statistical significance. We believe that the limited number of images that make up the dataset determine this discouraging result. Given the good results obtained, shown in the two tables 5.2 and 5.3, it is considered strictly necessary repeat the experiments on a larger dataset. At the moment this is not possible because there are no other datasets containing both the binocular views and the fixation maps to be used as ground thruth. So the first step to actually be able to affirm that the presence of depth discontinuity changes and improves the prediction of the positions of the salient points in the visual scene, is to formulate a new dataset that contains at least a hundred pairs of stereo images.

Chapter 6

Conclusions

Visual saliency is defined as "the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention". This concept can, indeed, be extended to each one of the 5 human senses. Each time that we explore the world around us, we receive a huge number of sensory inputs, impossible to elaborate all real time with the relative low number of resources available to our brain. For that reason our brain has developed the *selective attention* mechanism according to which, after a rapid scan of all the inputs, only a small portion, the most *salient* one, is selected for a more in-depth analysis. This *selective attention* mechanism involve, among all the other, the sense of sight. Studying its principles means understanding the methods our brain uses to select data to deepen and those to discard because cataloged as less salient after a first very quick analysis. This has been for decades field of research of important teams from Johns Hopkins University and from the ETH of Zurich, just to name two of the most important ones, that are interested in exploiting these principles in robotic applications. Nowadays the study of such mechanism is involving other fields such as advertising graphics. To approach this study, research has to be divided in more levels. It is needed to understand how the visual scene is organized and how the information are read and interpreted by the human visual system, this point is crucially important when, as in our case, the aim is to create a computational model of visual saliency as much biologically plausible as possible.

In the last five years the widely spread theory, according to which object features as color, intensity and orientation are perceived before the object itself (the so called Feature Integration Theory (FIT)), has been joined by the eve-increasing belief that features are nothing more the vehicle to divide the scene in a series of perceptual objects. The fundamentals of this statement are found in well-established studies as Gestalt psychology born around 1980 and the more recent Rensink theory. They are both based on the idea that the whole is different from the sum of its parts so it is not correct to say that perceiving features first is a fundamental step to recognized objects later, at the contrary the simple fact that an object exist can

be the reason why we perceive it, it exists as a super-ordinate phenomenon. These theories perfectly interact with the concepts of *Border Ownership* cells and *Grouping mechanism*, that suppose the existence, confirmed by neuronal activity registration experiments, of two type of neurons that work together to let us perceive the presence of a possible object within the first $50ms$ from the presentation of the stimulus. On these hypotheses the first biologically plausible computational model of proto-object based visual saliency is based. This model, proposed in 2012 by Russel et al. [28] breaks into a state of the art where more than half proposed models are based on the FIT principles. The performances of the model exceed those of the most widespread FIT-based models, confirming the now almost certain hypothesis that the decomposition of the visual scene into perceptual structures is a valid theory in the field of visual perception. This model does not deal with object recognition that is a mechanism that biologically occurs $120-150ms$ after the stimulus presentation, at the contrary it explains an alternative way to direct attention within the first tens of millisecond.

Starting from these good results, our study aims to add to the Russel et al. model the contribution of depth information to see if and how they change saliency areas prediction operated by the proto-object based model. Depth scene perception is something that we experiment every day but that, even if it has been demonstrated that the objects distribution in the space is decisive for the most salient areas selection, has been having a very marginal role in saliency models literature. We assumed that, as for two-dimensional features (color, intensity, etc.), depth discontinuity attracts human attention more than zones characterized by a constant depth and that grouping mechanism remain valid in depth perception theory. Differently from previous three-dimensional saliency models, ours does not required a pre-computed saliency map as input but directly uses stereopsis principles, inspired by human visual system. This latter is able to positioned an object in the space, starting from the relatives positions of its projection on the two retinas (left and right). This information is encoded by *Binocular cells*, neurons responsible for the perception of space according to planes parallel to the coronal head plane. Our algorithm identifies these plans starting from the analysis of a pair of stereoscopic images that are identical except for the fact that the position of the same image point is horizontally shifted between the two views. The different entity of this shift makes it possible to identify parallel planes at different distance from the observer, in three-dimensional space, just as it happens with the coding of the position of the projections on the two retinas. After having identifying the parallel planes, we suppose the grouping mechanism is activated within each plane. It is already known that grouping works at one feature level at the time: a proto-objects can be perceived because it lies on a unique depth plane as when it has a unique uniform color. Following these hypothesis we added to the previous version of the model [28] a new channel working alongside the already included channels related to two-dimensional

features and we developed the *Computational Stereo-Vision Model of Proto-Object based Saliency in Three-Dimensional Space* whose main steps are shown in figures 3.1.

After evaluating the correct functioning of the new channel (the *Disparity Channel*) using artificially created images were the where the shifts between th views were known, we tested the model on real scene images. From the comparison between three different evaluation metrics scores we demonstrate that the algorithm extension improves the model ability od predicting saliency areas: output saliency maps are more similar to the fixation density map included in the dataset and used as ground truth. Scores are visible in table 5.2 and table 5.3. This allows us to affirm that depth discontinuities influence the selection of the most salient areas within the image and that the grouping mechanism can be used to identify salient proto-objects even when working with three-dimensional perception.

Even if as the algorithm is know implemented it requires an estimation of the horizontal shifts that we expect to find between the tho views, information that should be given by the dataset used to test the algorithm itself, the overall idea is to contribute to te creation of a final computational stereo-vision model that will be able to automatically choose the range of disparity (horizontal shifts) to evaluate. It will follow the limit imposed by the resolution and the working method of far, near and tuned zero binocular cells of the visual cortex. Working on new channels that include new information, both two-dimensional and three-dimensional, on the visual scene, a higher output saliency map completeness level can certainly be reached.

A practical difficult with this approach is the paucity of dataset with bot stereoscopic views and a valid ground truth (i.e. fixation density map estimated using and eye-tracker) to evaluate the final saliency maps. The only complete dataset available online is composed by 18 images that are not enough to let our results to reach statistical significance. So to complete the study and find out if the information on depth discontinuity really improves the performance of the overall model, the first fundamental step is to create a new set of complete datasets.

Bibliography

[1] PHILOSOPHICAL TRANSACTIONS B, ed. *How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling*. 2017. URL: http://dx.doi.org/10.1098/rstb.2016.0113.

[2] Alexandra Battaglia-Mayer et al. *FISIOLOGIA MEDICA, 2nd edition*. Milano, Italy: Edi.Ermes srl, 2010.

[3] Kwabena Bohaen. "Verso l'occhio artificiale". In: *Le Scienze* (2005).

[4] Zoya Bylinskii et al. *What do different evaluation metrics tell us about saliency models?* Online from 5th of April 2017. URL: https://arxiv.org/pdf/1604.03605.pdf.

[5] Simone Campora. *Correzioni di Prospettiva*. Online, visited on 11th of November 2018. URL: http://www.simonecampora.com/blog/wp-content/uploads/2009/10/correzioni-di-prospettiva.pdf.

[6] Leonardo Chelazzi et al. "Rewards teach visual selective attention". In: *Vision Research* (2013).

[7] E. Craft et al. "A neural model of figure-ground organization". In: *Neurophysioly* (2007).

[8] Robert Desimone and John Duncan. "Neural Mechanism of Selective Visual Attention". In: *VISUAL ATTENTION* (1995).

[9] John Duncan. "Selective attention and the Organization of Visual Information". In: *Journal of Experimental Psychology:General* (1984).

[10] Simone Frintop, Erich Rome, and Henrik I. Christensen. "Computational visual attention systems and their cognitive foundations: A survey". In: *Cover Image* (2010).

[11] Simone Frintrop. *Computer Analysis of Human Behavior*. London: Springer, 2011.

[12] Morita Hiromi and Kumada Takasume. "Effects of pictorially-defined surfaces on visual search". In: *Vision Search* (2003).

[13] Brian Hu, Ralinkae Kane-Jackson, and Ernst Niebur. "A proto-object based saliency model in three-dimensional space". In: *Vision Research* (2016).

[14] Scientific & Medical ART (SMART) Imagebase, ed. *Coronal Plane Section of Head*. URL: https://ebsco.smartimagebase.com/view-item?ItemID=7101.

[15] L. Itti, C Koch, and E Niebur. "A model of saliency-based visual attention for rapid scene analysis"". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1998).

[16] team IVC and IRCCYN. *Gaze 3D dataset*. Online, downloaded on February 2018. URL: http://ivc.univ-nantes.fr/en/databases/3D_Gaze/?article1102.

[17] T. Judd, Durand F., and A. Toralba. "A Benchmark of Computational Models of Saliency to Predict Human Fixations". In: *Computer Science and Artificially Intelligence Laboratory Technical Report* (2013).

[18] C Koch and S Ullman. "Shift in selective visual attention: towards the underlying neural circuitry". In: *Human Neurobiology* (1985).

[19] C. Koch and S. Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry". In: *Human Neurobiology* (1985).

[20] Bruce M. Koeppen and Bruce A. Stanton. *Berne & Levy Physiology, 7th edition*. Philadelphia, PA, USA: Mosby Elsevier, 2018.

[21] Kurt Koffka. *Principles of Gestalt Psycology*. New York: Harcourt, Brace and company, 1935.

[22] Chiara Della Libera. *La percezione dello spazio e della profondità*. Online, visited on 11th of October 2018. URL: http://www.di.univr.it/documenti/OccorrenzaIns/matdid/matdid534706.pdf.

[23] D. Parkhurst, K. Law, and E. Niebur. "Modeling the role of salience in the allocation of overt visual attention". In: *Vision Research* (2002).

[24] Michael I. Posner and Steven E. Petersen. "The Attention System of the Human Brain". In: *Annual Review of Neuroscience* (1990).

[25] Fangtu T. Qiu and Rüdiger von der Heyt. "Figure and Ground in the Visual Cortex: V2 Combines Stereoscopic Cues with Gestalt Rules". In: *Neuron* (2005).

[26] Ronald A. Rensink. "The Dynamic Representation of Scenes". In: *VISUAL COGNITION* (2000).

[27] Alexander F. Russel. "Biofidelic Proto-Object Based Visual Saliency". Dissertation in conformity with the requirements for the degree of Doctor of Philosophy. Baltimore, MD, USA: The Johns Hopkins University, 2012. Chap. 2.

[28] Alexander F. Russell et al. "A model of proto-object based saliency". In: *Vision Research* (2014).

[29] D. Scharstein and C. Pal. "Make3D: Learning 3D Scene Structure from a Single Still Image". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2007).

[30] Scholarpedia, ed. *Border-ownership coding*. 2013. URL: http://www.scholarpedia.org/article/Border-ownership_coding.

[31] Gordon L. Shulman, Roger W. Remington, and John P. McLean. "Moving attention through visual space". In: *Journal of Experimental Psychology: Human Perception and Performance* (1979).

[32] Y. Sun and R. Fisher. "Object based visual attention for computer vision". In: *Artificial Intelligence* (2003).

[33] Y. Sun et al. "A computer vision model for visual-object-based attention and eye movements". In: *Computer Vision and Image Understanding* (2008).

[34] tobiipro, ed. *The human eye*. URL: https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/the-human-eye/.

[35] Anne M. Treisman and Garry Gelade. "A Feauture-Integretion-Theory of Attention". In: *COGNITIVE PSYCOLOGY* (1980).

[36] Matthieu Urvoy et al. "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences". In: *2012 Fourth International Workshop on Quality of Multimedia Experience* (2012).

[37] Dirk Walther, K. Law, and Christof Koch. "Modeling attention to salient proto-objects". In: *Neural Networks* (2006).

[38] Wikipedia. *Binocular neurons*. Online, visited on 25th of October 2018. URL: https://en.wikipedia.org/wiki/Binocular_neurons.

[39] Wikipedia. *Depth perception*. Online, visited on 8th of October 2018. URL: https://en.wikipedia.org/wiki/Depth_perception.

[40] Wikipedia. *Feature integration theory*. Online, visited on 28th of October 2018. URL: https://en.wikipedia.org/wiki/Feature_integration_theory.

[41] Wikipedia. *Gestalt psycology*. Online, visited on 28th of October 2018. URL: https://it.wikipedia.org/wiki/Gestalt_psycology.

[42] Wikipedia. *Image Rectification*. Online, visited on 11th of November 2018. URL: https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Image_rectification.html.

[43] Wikipedia. *Neuromorphic engineering*. Online, visited on 25th of September 2018. URL: https://en.wikipedia.org/wiki/Neuromorphic_engineering.

[44] Wikipedia. *Pyramid (image processing)*. Online, visited on 4th of October 2018. URL: https://en.wikipedia.org/wiki/Pyramid_(image_processing).

[45] UN. Wilming et al. “Measures and limits of models of fixation selection”. In: *PLoS ONE* (2011).

[46] J. He Zinjang and Ken Nakayama. “Visual attention to surfaces in three-dimensional space”. In: *Proceedings of the National Academy of Sciences of the United States of America* (1995).