Contribution to Perception and Artificial Bio-inspired Visual Attention for Acquisition and Conceptualization of Knowledge in Autonomous Robotics

Viachaslau Kachurka

UNIVERSITÉ PARIS-EST

Thesis

presented to obtain the title of
DOCTOR OF UNIVERSITY PARIS-EST in
Signals, Images and Automatics

by **Viachaslau KACHURKA**
viachaslau.kachurka@univ-paris-est.fr

Contribution to Perception and Artificial Bio-inspired Visual Attention for Acquisition and Conceptualization of Knowledge in Autonomous Robotics

Defended on 20/12/2017 in public and in presence of commission composed by:

Rapporteurs	Samia BOUCHAFA	Université d'Évry Val-d'Essonne
	Viktor KRASNOPROSHIN	Belarusian State University
Examinateurs	Christophe SABOURIN	Université Paris-Est
	Gilles BERNARD	Université Paris 8
Co-directeurs de thèse	Kurosh MADANI	Université Paris-Est
	Vladimir GOLOVKO	Brest State Technical University

UNIVERSITÉ PARIS-EST

Thése

présentée pour l'obtention du titre de
DOCTEUR DE L'UNIVERSITÉ PARIS EST en
Signal, Image, et Automatique

par **Viachaslau KACHURKA**
viachaslau.kachurka@univ-paris-est.fr

Contribution á la Perception et l'Attention Visuelle Artificielle Bio-inspirée pour Acquisition et Conceptualisation de la Connaissance en Robotique Autonome

Soutenue publiquement le 20/12/2017 devant le jury composé de :

Rapporteurs	Samia BOUCHAFA	Université d'Évry Val-d'Essonne
	Viktor KRASNOPROSHIN	Belarusian State University
Examinateurs	Christophe SABOURIN	Université Paris-Est
	Gilles BERNARD	Université Paris 8
Co-directeurs de thèse	Kurosh MADANI	Université Paris-Est
	Vladimir GOLOVKO	Brest State Technical University

Abstract

Dealing with the field of "Bio-inspired Perception", the present thesis focuses more particularly on Artificial Visual Attention and Visual Saliency. A concept of Artificial Visual Attention, inspired from the human mechanisms, providing a model of such artificial bio-inspired attention, was developed, implemented and tested in the context of autonomous robotics. Although there are several models of visual saliency, in terms of contrast and cognition, there is no hybrid model integrating both mechanisms of attention: the visual aspect and the cognitive aspect.

To carryout such a model, we have explored existing approaches in the field of visual attention, as well as several approaches and paradigms in related fields (such as object recognition, artificial learning, classification, etc.).

A functional architecture of a hybrid visual attention system, combining principles and mechanisms derived from human visual attention with computational and algorithmic methods, was implemented, explained and detailed.

Another major contribution of this doctoral work is the theoretical modeling, development and practical application of the aforementioned Bio-inspired Visual Attention model, providing a basis for the autonomy of assistance-robotic systems.

The carried out studies and experimental validation of the proposed models confirmed the relevance of the proposed approach in increasing the autonomy of robotic systems within a real environment.

Keywords : Artificial visual attention, soft computing, bio-inspired perception, genetic algorithms, artificial vision, visual saliency, robotics

Résumé

La présente thèse du domaine de la “Perception Bio-inspirée” se focalise plus particulièrement sur l’Attention Visuelle Artificielle et la Saillance Visuelle. Un concept de l’Attention Visuelle Artificielle inspiré des êtres vivants, conduisant un modèle d’une telle attention artificielle bio-inspirée, a été élaboré, mis en œuvre et testé dans le contexte de la robotique autonome. En effet, bien qu’il existe plusieurs dizaines de modèles de la saillance visuelle, à la fois en termes de contraste et de cognition, il n’existe pas de modèle hybridant les deux mécanismes d’attention : l’aspect visuel et l’aspect cognitif.

Pour créer un tel modèle, nous avons exploré les approches existantes dans le domaine de l’attention visuelle, ainsi que plusieurs approches et paradigmes relevant des domaines connexes (tels que la reconnaissance d’objets, apprentissage artificiel, classification, etc.).

Une architecture fonctionnelle d’un système d’attention visuelle hybride, combinant des principes et des mécanismes issus de l’attention visuelle humaine avec des méthodes calculatoires et algorithmiques, a été mise en œuvre, expliquée et détaillée.

Une autre contribution majeure du présent travail doctoral est la modélisation théorique, le développement et l’application pratique du modèle d’Attention Visuelle bio-inspiré précité, pouvant constituer un socle pour l’autonomie des systèmes robotisés d’assistance.

Les études menées ont conclu à la validation expérimentale des modèles proposés, confirmant la pertinence de l’approche proposée dans l’accroissement de l’autonomie des systèmes robotisés – et ceci dans un environnement réel.

Mots clés : Attention visuelle artificielle, soft computing, perception bio-inspirée, algorithmes génétiques, vision artificielle, saillance visuelle, robotique

Acknowledgements

First of all, I would like to express a feeling of my deepest gratitude to the directors of my thesis work – Prof. Kurosh Madani of Université Paris-Est Creteil and Prof. Vladimir Golovko of Brest State Technical University. It was their colossal effort and collaboration which made this work possible in first place. And it is not only their scientific experience and aide in research, not only their patience in superivision of my PhD studies, but also the unending support in many different aspects of a student's life between two countries and two universities.

Also, I would like to voice my thanks to Dr. Christophe Sabourin of Université Paris-Est Creteil, who's ideas and remarks helped greatly in the process of this work.

I want to thank deeply Prof. Uladzimir Rubanau of Brest State Technical University, who had always supported this work, always had faith in me and helped to get through millions of different administrative problems of co-supervised PhD study.

I want to thank Prof. Victor Krasnoproshin from Belarusian State University, Prof. Samia Bouchafa from University Évry Val d'Essonne, and Prof. Gilles Bernard from University Paris 8 for having accepted to be my PhD thesis referees.

I would like to thank the members and ex-members of LISSI laboratory, including Dr. Aurélien Hazan, Dr. Amine Chohra, Hossam Fraihat, Yu Su and Roozbeh Sadeghian. I would also like to thank my parents, Dr. Anatoli Kachurka and Iryna Kachurka, my grandmother Evdokiya Baunina, my brother Dr. Pavel Kachurka and his wife Volha Kachurka for ever-lasting support and help during this research, as well as my friends Yulia Davidiuk, Siarhei Shpak, Nikita Lepeiko, Katerine Shtamburg and many others. Special thanks goes to my 3-year-old nephew Yaraslau Kachurka, who inspired some experiments done during this study.

Contents

List of Figures

List of Tables

General Introduction

Foreword

In the last decades robotics was developing intensely and productively in different niches: industry, engineering, aviation, and even extreme activities (like exploration in hazardous conditions) — the robots took and hold their places in working process firmly. However, all these types of robots in most cases have always been highly specialized and thus not expected to have capabilities for general "human–robot" interactions (as it is usually a secondary or even non-important task).

Yet, this tendency have begun to change in the last decade, as the market of "personal robots", – robots, which are expected to co-exist with humans in daily life and interact with them closely, – started to emerge. Thus, in 2007 Bill Gates predicted [Gates 07] exponential market growth relating personal robots. In 2012 valuation of the market of personal robots was assessed as $ 1.6 billion, and in 2017 is expected to increase 4 times — up to $ 6.5 billion [Simon 14]. For example, in 2014 investments of $ 2 million, targeted onto development of a simple personal robot *JIBO*, were gathered by crowdfunding during just one day [ABIResearch 13], and such facts let us consider this market having stable and high level of interest of customers and potentially high profitability of any start-up in this field.

But in past few years the industry of personal robots have shown the imperfection of software, which also did not let the robot look and feel compared to human, accentuating the problem of uncanny valley [Mathur 16]. All this consideration can also be implied on the area of "social robots", – robots that interact and communicate with humans by following social behaviors and rules attached to its role.

Another field, which concerns both the personal and social robots, is the problem of machine vision — including the approach of robot vision system to human level of vision for higher rate of success in a number of tasks, such as robotic driver assistant system or the task of finding and recognizing faces [Fritsch 08, van Kleef 16]. However, these problems are also almost solved nowadays, due to being set within more narrow limits and constraints.

More general tasks, such as tasks of perception and recognition of environment

[Ittelson 76], with varied success were solved in computer and machine vision using limited approaches [Franke 05, Giovani 15].

A successfullness of "human-like approach" have been shown at first time by Laurent Itti in 1998 [Itti 98], incorporating the notions and concepts of neurobiology and psychology of primates into the field of computer and machine vision de-facto, linking this field to biological patterns of visual perception tasks solving. Following this research, in first two decades of XXI century, a large number of works was devoted to various aspects of such an approach. However, their overwhelming majority, due to high computational resource requirements, was still oriented onto solving narrow, limited tasks [Borji 13a]. As an example — search and recognition of text in the image [Lienhart 02, Erkan 04], recognition of car license plates [Chen 09]. This also left a mark on the possibility of research towards human-like approach in visual perception in machine vision and robotics, as the embedded CPUs of autonomous robots could not provide sufficient resources for image treatment in real time.

However, in last years several factors coincided: fast convolutional neural networks, usable for image processing (e.g., GoogLeNet [Szegedy 15]); new generation personal robots of the type "robot-companion" with sufficient computing resources (e.g., Aldebaran Pepper [Ebling 16]); stable growth of interest and demand on Asia markets. All this let us say about a possibility of research in terms of human-like approach on a new level. Thus, imitation of human gaze in robots can help, for example, to approach a solution of "uncanny valley problem" [Koschate 16] or visual help problem [Sudol 10].

Development of a human-like complex visual attention system based on artificial convolutional neural networks and human-like approach could augment quality of interaction between human and autonomous robot in general case, and also set foot onto approach to more efficient solution of the problems of visual search in different contexts (such as a search for fire residues [Toulouse 16] or navigation [Chang 10]).

Motivation and Objectives

As stated in Foreword, the problem of a human-like complex visual attention is one of the fields in the social robotics industry, – the niche which gains in investments each year.

As the science fiction authors' dreams stay far in future, small steps in this direction are done continuously. In fact, such steps could relate also the field of visual attention mechanism, which might approach by efficiency the human's visual attention. Yet, this mechanism is worked upon only partially, being a problem

in narrow contexts which is successfully solved by both bio-inspired and classical approaches.

As the pioneer of artificial intelligence, Roger Schank, said in the preface of his book "Dynamic Memory Revisited" [Schank 99, p. vii], *"In the ... 1980s I was fascinated by the idea that computers could be as intelligent, as people. ... I no longer hold such views"*, – scepticism about the "turtle" speed of general AI research as opposed to practical AI-based solutions, is pretty demotivating. Yet, any contribution to this domain adds a sand grain into the pile, which might one day finally become the Giza pyramid.

Another point of view on motivation can be found in opposition of terms "automatic" and "autonomous". For example, in more general context of machinery, "automation" can be defined as "...the technology by which a process or procedure is performed without human assistance" [Groover 07], while "autonomy" can be seen as " a possibility of choice to make free of outside influence" [Clough 02]. When we narrow the context to the robotics, the term "automatic" can be applied to industrial robotics in general [Shell 00], or to any programmable robot as it can implement the preexisting algorithm "automatically". Autonomous robot, on the other hand, can be seen as "intelligent machines capable of performing tasks in the world by themselves, without explicit human control" [Bekey 05].

Thus, contemporary machines are often automatic, but almost never fully autonomous. This also applies to the context of the visual attention – the biggest machine vision victory so far is face or human stature recognition, combined with simplistic gist approaches in form of scanning QR-codes or following the lines. This is why the concept of bio-inspired human-like machine visual attention, based on generalized approach, is so important for future systems and intelligent robots. It is because this is the way of a major contribution to a true autonomy of future intelligent systems, – including not far fetched social robots, or general humanoid robotics in future.

In accordance with the requirement of autonomy in context of the machine visual attention, we can set up the following objectives for the work that is developed throughout the present thesis:

- Explore the existing state of art in the field of visual attention, the problems posed and solved by the best approaches and models in general, in order to investigate the necessity of contribution to novelty models;

- Contribute to conception of a visual attention system, which is constructed by several basic principles of human's attention mechanism;

- Contribute to a model, which can provide autonomous visual attention decisions, thus placing decision-wise "autonomy" as a main skill with notion of possibility of real-time processing, and be able to show the efficiency not far worse than state of art visual attention models, capable of flexible tuning, and acting as a part of the previously mentioned visual attention system;
- Create a functional working implementation of this system, which should prove the mobile robotic platforms capable to solve different real world tasks.

Frame of the Work

This thesis is done as a co-supervised work between University Paris-Est and Brest State Technical University under co-direction of Prof. Kurosh Madani (LISSI - Laboratoire Images, Signaux et Systèmes Intelligents) and Prof. Vladimir Golovko (LANN - Laboratory of Artificial Neural Networks).

The work in the scope of this thesis is done in the form of constructive research, based on definition of the problem via state-of-art study in bio-inspired visual attention field and the application of approaches from that field to the autonomous robotics.

The theoretical body of knowledge for such research considers the corpus of previous works on the topic in LISSI team SYNAPSE (SYstèmes cogNitifs Artificiels et Perception bio-inSpiréE), – including [Ramík 11], [Amarger 12], [Wang 12], [Madani 12] and [Ramík 13], – thus aiming the thesis to continue the research relying on several attention concepts already outlined in [Ramík 12] with inclusion of several decision-related artificial intelligence aspects based on LANN team previous works [Golovko 03], [Imada 07], [Kachurka 12].

The already existing context of artificial curiosity, feature extraction-based visual recognition, visual saliency concept and knowledge extraction have incepted the frame of this work, which is focused on approaching a human-level skill of perception as a property of artificial visual attention in order to improve the level of decision autonomy.

The theoretical framework of this thesis employs several methodological tools, such as simulation and modeling methods, widely used in the field of visual attention in order to imitate the complex procedure of human psycho-physiology of acquiring and processing visual information and to test these imitation models on synthetic simulations of such acquisition; the field of autonomous robotics infers also employment of experiment-based method, in order to evaluate and validate the implementations of theoretical models by running them on programmable robotic

platforms.

Contribution

The work accomplished in this thesis resulted in several contributions, relating artificial visual attention systems:

- First, the compilation study of state-of-the-art on visual attention field, which allows objective evaluation of achievements, relating autonomous visual attention in machines as well as an outline of open problems in this domain;
- The second major contribution of this thesis is the low-level attention model, serving as further improvement of previously created contrast-based saliency approach. This low-level attention model consists of several bio-inspired techniques (visual saliency, center-peripheral antagonism) and is able to learn by evolutionary algorithm approach in order to improve the quality of results in simulations and real-world experiments, also can be a key part in bigger visual attention model;
- Third is the contribution in the form outline of bio-inspired visual attention system ("combined model") which is based on both attention mechanisms – previously mentioned low-level features-based attention model as key part, basic level attention, and cognitive phenomena-based attention model as high level attention, – using also human-like working memory model as controlling module. This system stands out from similar existing algorithms as it has more complex, human-inspired structure and could be used either as standalone mechanism, or as basis for the further complexification by adding new components;
- The last major contribution is implementation of the combined model in mobile autonomous robotic platforms, such as WiFiBot-M, Aldebaran NAO and Aldebaran Pepper, along with validation on several real world tasks (fire detection, attention concentration, visual exploration) solved by the proposed system.

Thesis Organization

This thesis is constituted of four chapters, leading the reader from the state of the art through general concepts of the whole system, detailing of different parts,

to its concrete implementation and verification via several applications in real world environment.

Chapter 1 introduces the reader into the state of the art in the domain of visual attention. It discusses the domain itself and reflects it through model-based methodology as the classification of problems and models in the field, along with the means of their evaluation and benchmarking. Among others it discusses works in "bottom-up" saliency, solving the eye fixation problem, which then is posed under different angles throughout the whole thesis. After the discussion on "top-down" saliency, the chapter also briefly considers several related fields, such as visual object recognition, knowledge storage and lexical analysis. This chapter should give the reader a general overview on the state of the art, existing techniques and terminology on which we further develop our research described in chapters that follow.

In Chapter 2 the general scheme of the proposed combined visual attention model is outlined, defining constitutive parts of the model along with some general considerations and conditions met with the units of such a model, which are subsequently concretized in Chapter 3. It represents the concept of combined visual attention as union between two different approaches, moderated via a decision module.

Chapter 3 is focused on details for each of aforementioned model units; thus, it shows theoretical basis for the key part of the system – low-level "bottom-up" attention mechanism, which can also work independently. While first part of the chapter introduces several algorithmic steps borrowed from precious researches, second part introduces several bio-inspired techniques in order to improve the overall efficiency. Third part shows the evolutionary-based tuning process, along with the evaluation of its flexibility and ability to generalize, comparison of the algorithm against modern top algorithms on several benchmarks, as well as the time complexity of the algorithms, and a novel mean for evaluation and saccade modeling. Then the chapter explores the informed choice for several techniques in the field of visual object recognition in order to constitute second part of combined model, – the "top-down" attention mechanism. After all the chapter is concluded with detailed description of third part of the combined model, – the "moderating" decision module, – as well as detailed scheme of the whole model and generalized algorithm of operation.

And the Chapter 4 finalizes the research by presenting the combined model in the form of its implementation on several autonomous robotic mobile platforms, as well as addressing some implementation-wise quirks. Also it shows several experiments of the real world indoor environment, as well as a chaotic outdoor environment.

The closing chapter of this thesis is the General Conclusion. That is where the reader is given a summary conclusion and an evaluation of the research presented here. Finally, perspectives of possible future directions of the work are provided.

1 | The State-of-art Relating Visual Attention and Perception

1.1 Introduction

Human-like visual perception is based on one of the important concepts in primate neurobiology – so called "saliency" of an object, which is directly linked to the concept of visual attention. Evaluating these concepts as measurable characteristics of different parts of image provides practical approaches to solve different types of problems mostly from the field of computer and machine vision, as well as if applied to robotics [Borji 10, Ouerhani 05, Scheier 97, Courty 03].

In this chapter we consider existing and state-of-art approaches to visual perception in computer vision, evaluate their efficiency and caveats. Also we define more precisely the tasks of this research, input data used for analysis, and quality parameters used for evaluation.

1.2 Problem of visual attention in computer vision

Let us start with the main concepts, terms and definitions, given and used in this work. As it is a work based on concepts, taken from intersection of neurobiology, computer vision and psychology, we will need to provide the definitions as given in well-known reputed sources in these fields.

Attention *is the behavioral and cognitive process of selectively concentrating on a discrete aspect of information, whether deemed subjective or objective, while ignoring other perceivable information.* [Anderson 90]

Gaze *is a coordinated motion of the eyes and head, which has often been used as a proxy for attention in natural behavior.* [Hayhoe 05]

Thus ***visual attention*** can be defined as *one's perception of one of the aspects of information via visual sensors, shown by gaze or its imitation, and one's*

concentration on on it.

In the field of computer vision the task of modeling human's visual attention creates another term – "saliency", often interchangeable with "visual attention", whilst being slightly different. As it is given, for example, in one of the key publications in the field:

Saliency *intuitively characterizes some parts of a scene – which could be objects or regions – that appear to an observer to stand out relative to their neighboring parts.* [Borji 13a]

At the same time attention is a general concept, covering all factors that influence selection mechanisms. Thus we can say that saliency can be interpreted as an intuitive characteristic, usable in visual attention model in order to describe it.

In general, there is no absolute measure for saliency; it is always used as a relative metric. If some part of visual scene is more salient, than another part, it means that this part is more important at this exact moment.

We need also to underline, that the word "model", as seen in most papers in the field, such as the best-known [Itti 98], is mostly used in the sense of simplistic operational model of human visual attention, where the process of acquiring and processing of visual input by physical and psycho-neurological means of a human is imitated by a set of procedures or algorithms, whether they are bio-inspired or purely mathematical. Therefore throughout the scope of this thesis we will also use this word in the same way, as a soft synonym for "algorithm" and in the sense of modeling the means of human's visual attention as a set of procedures.

1.2.1 Visual attention problem

Let us define a visual attention problem as a problem in computer or machine vision, where the solution should make usage of a visual attention model, and the input is given in form of an image or a video. In general, a visual attention problem may be formulated as following: *an input image is given as a set of pixels, each represented by a 2-dimensional position and color (given in any image space); using this information, we have to estimate an attention characteristic for each pixel.* In many cases input image may be replaced by series of images, which could be treated as video frames.

There exist several ways of classification for such problems. Thus, [Borji 13a] categorizes these models using 13 factors, ordered by priority and generalization.

First two factors are related to the direction of information processing and might be looked upon as the main factors [van de Weijer 04], as shown on Figure 1.1:

- "top-down" factor, where visual scene perception is defined by viewer's cogni-

tive phenomena: knowledge, experience, reward or target in each moment of time. A classic example is given in [Yarbus 67], where the gaze of a subject human is differently perceptive for the same visual scene, depending on the target given by examiner.[1]

- "bottom-up" factor, where visual scene perception starts directly from its characteristics (another name for such direction is "stimulus-driven"). Such factor was firstly shown in [Treisman 80], when subject's gaze was catching the only horizontal line among all the vertical ones;
- combination of "top-down" and "bottom-up" factors.

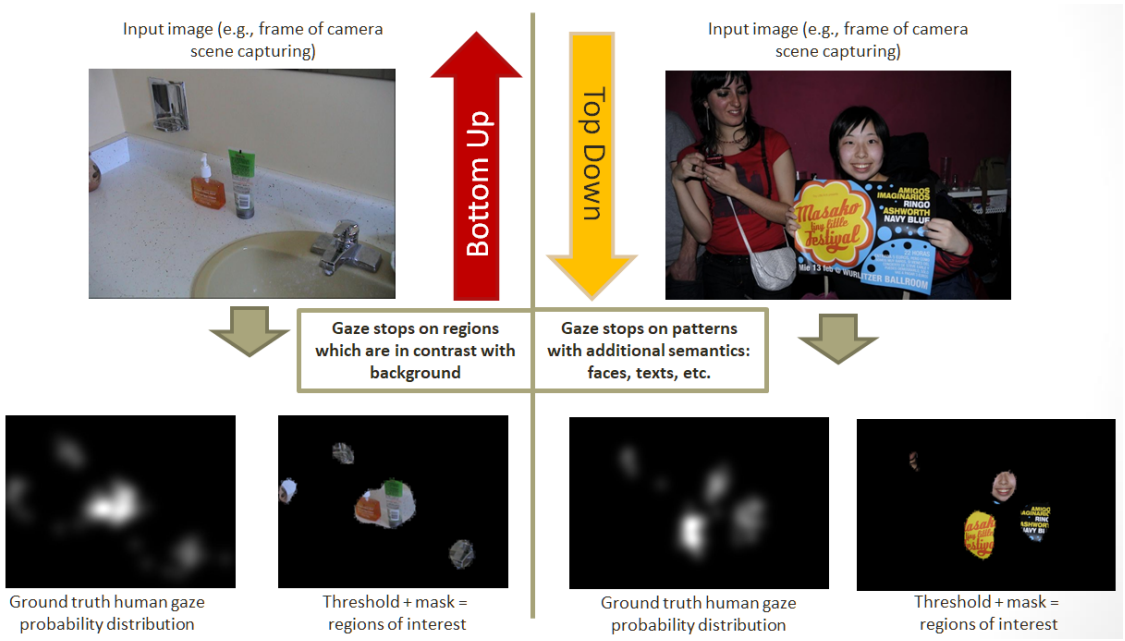

Figure 1.1: Examples of two different reasonings for bottom-up and top-down directions of attention, where the first tends to apply itself to contrast objects, the second is applied to cognitive phenomena such as faces or texts

In general, a visual attention model should combine both factors to be able to stand for real human visual attention. But practically the implementation of such model is difficult due to computational problems linked to "top-down" direction. For this reason in such models some a priori knowledge of the problem being solved is often used to narrow down the usage and computational complexity – such as an assumption of the placement of object in question in the problem of license plate recognition, or knowledge about human face form [Goferman 12, Cerf 08].

[1] More on the topic of "top-down" and "bottom-up" attention can be found in psychoneurological reviews, e.g., [Hayhoe 05] or [Triesch 03]

Most of the models, which use the "bottom-up" direction, use the previously given interpretation of "relative saliency" and implement it into "saliency maps", firstly coined by Koch and Ullman in [Koch 87] as a variety of a heat map:

Saliency map *is such map, where for each individual point there exists an estimation of its saliency value, relative to all other separate points of this map.*

Thus, saliency map mechanism usage suggests existence of an algorithm, which creates such an estimation of relative saliency value for each point in question. Most of the models, which use saliency map mechanism, are built around such algorithms. The saliency maps, created by such algorithms, are usually represented as images of the same size and form as input visual scenes provided and encoded as images, where each pixel of saliency map image contains information of relative saliency for the same-position pixel of the image in question, thus implementing the aforementioned point-to-value correspondence. Two most popular mechanisms of encoding in this case are heat map and the shades of gray (example is shown in Figure 1.2).

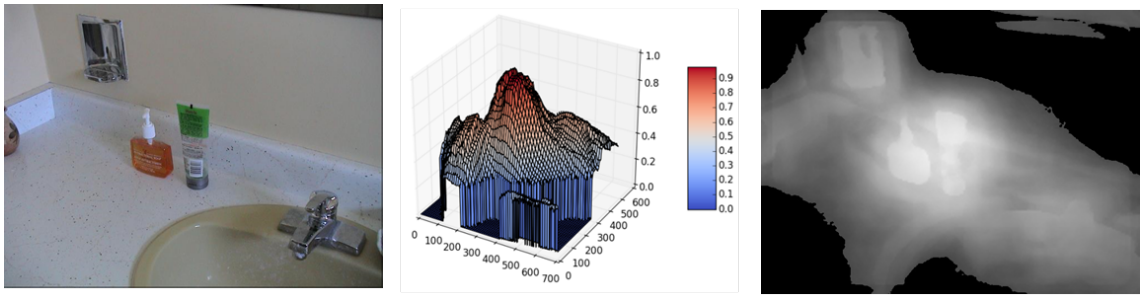

Figure 1.2: Example of input image, which after processing gives a prediction distribution, usually encoded as grayscale map

Another approach to model classification is based on categorization of the problems being solved [Borji 15]:

- A problem of predicting the gaze fixation, also called "eye-fixation problem", where the saliency map is usually interpreted as the map of distribution of human's gaze fixation probability for each independent point of the map. Such problems usually treat the saliency map as the sought solution (examples could be found in [Wolfe 05, Parkhurst 02, Li 10, Borji 12b, Borji 12a, Koehler 14, Li 14]).

- A problem of identification of salient regions or objects in the scene, where the estimation of saliency map is usually only one of the steps in the algorithm, and the sought solution should be given in the form of a part of the initial visual

scene, which represents the most salient region or object (e.g., [Achanta 09, Tian 15, Wang 06, Borji 13b]).

Although there is an apparent affinity between these two types of problems, implicitly hinting onto a possible interchangeability of the saliency maps between models of these two types, these saliency maps often are vastly different due to the difference of the problems being solved. Though common approaches to saliency map calculation are usually the same, the difference is in the model details. Consequently, the models used to solve the problems of these two types, could also be classified by the type of the problem.

Eye fixation prediction problem. Eye movements are often treated as external evidence of attention shift. For example in the experiments, where the stimuli on the scene are not very well seen, fixation on each of them takes longer time while the saccadic movements are shorter [Rayner 98].

Saccade, *or* ***saccadic movement,*** *is a quick, simultaneous movement of both eyes between two or more phases of fixation in the same direction.* [Javal 78]

Saccadic movements are ballistic in their nature. After such movement starts, it will end independently of changes in the ending fixation point. This means that each new saccadic movement is defined fully in advance, before it starts. Such small paradox brings us to another concept as given in [Wong 81]:

Saccadic programming *is a process of saccadic direction and maintenance, which defines each next movement in advance; different parts of human nervous system participate in this process (cranial nerves, cores of midbrain tegmentum, etc.)*

Thus the task of prediction of eye fixation can be interpreted as a modeling problem for the behavior of the corresponding elements of human's nervous system, participating in saccadic programming. Such models can be called as saccadic programming models.

Such models could be useful for different image processing tasks, where information amount should be reduced – scaling, thumbnailing, – or in marketing tasks.

Salient regions identification problem is also known as "salient object detection problem", and is often treated as a problem of search and extraction of one or several salient objects out of the image. Main difference from the first type problems in this case is that it not enough only to predict all the main points of eye fixation; we have to as correctly as possible find all the points depicting the object (or the part of the object) to provide them to the next steps of algorithm. Most often such

step is "extraction" – clipping of the image or blackening the non-salient points. In this case the necessity of object search adds to the model the necessity of additional operations on the image. Usually such operation is either ***segmentation*** or ***edge detection***, which allows us to estimate the borders of objects or object-like regions of the image.

The solution model for the object extraction problem can be used in many tasks of computer or machine vision – classification, recognition, tracking, search, etc. [Borji 13a]

1.2.2 Bottom-up visual attention direction

To solve a problem of the first type (eye fixation problem), or the main part of the problem of second type – saliency map estimation to predict the eye fixation, – most often one of the three general approaches is used to create a bottom-up model.

Classic contrast-based approach have been firstly coined by Koch et Ullman in 1987 [Koch 87], then it was "rethought" and redefined many times by researchers in their works (e.g., [Itti 98, Parkhurst 02, Wolfe 05, Achanta 09, Li 10, Liu 11, Borji 12b, Cheng 15]) – models which use this approach with some changes. In general, this approach is based on hypothesis that each point can be seen as more salient if it has more contrast in relation to other points of the image. Thus is implied a relation if not of equivalence, than at least of dependency between "contrast" and "saliency", and also between their abstract quantitative estimations. But the term of "contrast" in this context is not detailed or concrete, and is given to be completed by researchers for each model.

Thus, the most well known interpretations of this approach were given by Itti in [Itti 98] and developed by Achanta in [Achanta 09] – where the quantitative estimation of contrast is defined as difference of some quantitative characteristics of each point from the mean characteristics of the whole image. Thus, if a "point" is treated as a pixel in this context, it is defined by three quantitative characteristics – colors in 24-bit RGB color system, – and a relative estimation of contrast can be defined as a sum of differences of each color from the mean image colors.

Another version of interpretation was explored by Liu in [Liu 11]. Based on the concept of "central-peripheral antagonism" [Mach 65, Hering 74, Westheimer 04], which implies that the center and the margins of mammal's retina show antagonistic behavior to one another, it was proposed to use so called "local contrast", where an estimation for the contrast of each point is given as a difference between mean quantitative characteristics for near-neighbouring ("center") and little-more-

far-neighbouring ("periphery") points of the image.

Both main variations of this approach are based on pretty simple linear calculations of statistical indicators, and this makes the approach pretty simple and non-demanding in terms of computational resources. On the other hand, this approach gives worse performance evaluations in comparison with another two approaches. Thus such an approach is usually used today only in the cases when the problem of limited resources is important enough and with higher priority than the problem of more accurate result.

Feature extraction-based approach is a group of models, which try to find individual elements of the image, and characteristics of these elements. This approach tries to distance itself from the search of "contrast", although sometimes in some models it can incorporate such a step. Also due to its nature it can be seen as an ensemble of completely different interpretations. A feature in this context is interpreted as any characteristic: orientation, form, relative size of the element [Zhang 13, Borji 12b]; direction of movement or amplitude of repeated movements in video [Cassagne 15]; any feature, which can be seen as typical or descriptive for given problem context.

Such definition allows interpretation of many feature-based models as the models with factor combination of bottom-up and top-down directions of attention.

Among the big set of interpretation two most popular could be distinguished:

- "Rarity of features" [Borji 12b, Riche 13a, Riche 13b, Cassagne 15] – if any feature is shown on the image more rarely, than other comparable features, than the element containing such a feature would be more salient. As an example for such approach is usually given a synthetic picture, where among many small rectangles there are several same-sized figures of a different form (for example, circles). In this case exactly due to the rarity of the feature "circle form" exactly these figures would be salient.

- "Complex edge search" [Borji 15, Zhang 13] – the edges of objects are defined as the features to be searched, and they are treated as elements with a feature "line complexity". In this case the edge is interpreted as a line, consisting of an ensemble of high-order curves. A synthetic characteristic is defined, directly dependent on the number and degrees of the curves. In this case an element of the image with the highest estimation of this characteristic, which could be interpreted as "element with the most complex edge", is usually thought to be the most salient one.

In practice this approach is usually combined with searching of another feature "relatively small size", because an object with a complex form, occupying a significant (yet not big) part of the image, will be more salient, than too small or too big object.

Such approach in any of its interpretations would usually need more computational resources, than the classic approach, while giving better performance.

Neural network approach started to be vastly developed in last several years because of development and evolution of convolutional deep neural networks. This approach might be based on one of two main principles. First is that convolutional neural networks are very good in terms of finding correlations and regularities both among evident and among hidden features of images. Second principle is that a deep neural network might be interpreted as imitation (in first approximation) of the multilayer system of saccade programming in human's neural system, which makes it more than just an abstract mathematical model.

Most of well-known models work using first principle [Judd 12, Vig 14, Liu 16, Kümmerer 16, Kruthiventi 17].

This approach gives best results in comparison to other approaches, but requires much more resources for learning and operational cycle of the network. Thus, network **DeepGaze II** [Kümmerer 16], implemented as a web-service, even in times of minimal load onto the server, requires about two minutes for one image processing. From the point of view of autonomous robotics, the question of available resources might be the key part in decision process while choosing the approach. Due to this, the usage of classic contrast-search approach (possibly partly mixed with two others) we might consider as appropriate for models, designed to be executed in autonomous robots.

1.2.3 Top-down visual attention direction

As given in [Borji 13c], the top-down attention direction is based on cognitive phenomena, which are usually hardly formalized.

Models have explored three major sources of top-down influences in response to this question: *"How do we decide where to look?"* Some models address visual search, in which attention is drawn toward features of a target object we are looking for. Some other models investigate the role of scene context or gist to constrain locations that we look at. In some cases, it is hard to precisely say where or what we are looking at since a complex task governs eye fixations, for example, in driving. While, in principle, task demands on attention subsume the other two factors, in

practice models have been focusing on each of them separately. Scene layout has also been proposed as a source of top-down attention [Navalpakkam 05, Oliva 01] and is considered together with scene context.

Object Features. There is a considerable amount of evidence for target-driven attentional guidance in real-world search tasks [Einhäuser 08, Pomplun 06]. In classical search tasks, target features are a ubiquitous source of attention guidance [Zelinsky 08]. Consider a search over simple search arrays in which the target is a red item: attention is rapidly directed toward the red item in the scene. Compare this with a more complex target object, such as a pedestrian in a natural scene, where, although it is difficult to define the target, there are still some features (e.g., upright form, round head, and straight body) to direct visual attention [Ehinger 09]. The guided search theory [Wolfe 07] proposes that attention can be biased toward targets of interest by modulating the relative gains through which different features contribute to attention. To return to our prior example, when looking for a red object, a higher gain would be assigned to red color.

Scene Context. Following a brief presentation of an image (80 ms or less), an observer is able to report essential characteristics of a scene [Bailenson 05]. This very rough representation of a scene, so-called "gist", does not contain many details about individual objects, but can provide sufficient information for coarse scene discrimination (e.g., indoor versus outdoor).

It is important to note, that gist does not necessarily reveal the semantic category of a scene; [Chun 98] have shown that targets appearing in repeated configurations relative to some background (distractor) objects were detected more quickly [Joubert 08]. Semantic associations among objects in a scene (e.g., a computer is often placed on top of a desk) or contextual cues have also been shown to play a significant role in the guidance of eye movements [Hwang 11]. Several models for gist utilizing different types of low level features have been presented. [Oliva 01] computed the magnitude spectrum of a Windowed Fourier Transform over non overlapping windows in an image. They then applied principal component analysis (PCA) and independent component analysis (ICA) to reduce feature dimensions; in [Walker 02] the researchers applied Gabor filters to an input image and then extracted 100 universal textons selected from a training set using K-means clustering. Their gist vector was a histogram of these universal textons.

Task Demands. Task has a strong influence on deployment of attention [Yarbus 67]. It has been claimed that visual scenes are interpreted in a need-based manner to

serve task demands [Triesch 03]. [Hayhoe 05] showed that there is a strong relationship between visual cognition and eye movements when dealing with complex tasks; subjects performing a visually guided task were found to direct a majority of fixations toward task-relevant locations. It is often possible to infer the algorithm a subject has in mind from the pattern of her eye movements.

For example, in a "block-copying" task where subjects had to replicate an assemblage of elementary building blocks, the observers' algorithm for completing the task was revealed by patterns of eye movements. Subjects first selected a target block in the model to verify the block's position, then fixated the workspace to place the new block in the corresponding location [Ballard 07]. Other research has studied high-level accounts of gaze behaviour in natural environments for tasks such as sandwich making, driving, playing cricket, and walking (e.g., see [Henderson 99, Rensink 00, Bailenson 05]).

1.3 Model quality evaluation in visual attention

Comparison of quality (performance, efficiency) of how the models work could be done using several different parameters. For both types of tasks there exist different sets of evaluation protocols. Thus, the second task (salient regions identification problem) can be interpreted as a classification problem, where each element of the input visual scene (e.g., image pixel) could be classified as either "salient" or "non-salient". In this case model quality assessment is done using the approach adopted in the field [Powers 11]:

- ***FP*** *(false positives)* – number of non-salient elements, classified as salient, – also known as Type I errors;
- ***FN*** *(false negatives)* – number of non-found salient elements, – also known as Type II errors;
- ***TP*** *(true positives)* – number of salient elements, correctly classified;
- ***TN*** *(true negatives)* – number of non-salient elements, correctly classified.

Based on these classic metrics are another de-facto standards ***Precision / Recall / F-measure***:

- ***PR*** *(precision)* – precision of classification which shows the quality of how the system works in whole:

$$PR = \frac{TP}{TP + FP} \tag{1.1}$$

Table 1.1: Average metrics of the best models benchmarked over **MSRA10K** dataset

Algorithm	PR, %	RE, %	F, %
Borji's SLIC model [Borji 15]	87.2	72.1	83.5
Cheng's global contrast H-model [Cheng 15]	72.9	71.8	72.3
Cheng's global contrast R-model [Cheng 15]	83.1	58.8	76.5
Achanta's frequency contrast saliency [Achanta 09]	69.1	54.9	65.4
Goferman's context saliency [Goferman 12]	59.8	56.7	58.1

- ***RE*** *(recall)* – quality of detection of salient elements, also known in terms of ROC-analysis as sensitivity [Fawcett 06], – it shows the percent of elements classified as salient being really salient:

$$RE = \frac{TP}{TP + FN} \tag{1.2}$$

- ***F*** *(F-measure, also known as F1)* – combined metric which is calculated based on precision and recall:

$$F = 2 * \frac{PR * RE}{PR + RE} \tag{1.3}$$

To evaluate models, and have comparable metrics, we need also to define a single testing dataset; such a dataset in last years for the second type tasks is Zhang's **MSRA10K** dataset [Zhang 13].

In Table 1.1 we can see average metrics of several models, which try to solve the problem of extraction of salient objects from visual scenes, represented by single-standing static photos from **MSRA10K** dataset, consisting of 10 thousands testing images. All these models represent the "bottom-up" approach, as they are essentially based calculation-wise on one or several low-level features of the input images.

Eye fixation problem, on the other hand, can not be interpreted directly as a classification problem. Model quality is defined by comparison of two saliency maps, – calculated map and "ground truth" map, – where an empiric map of distribution of eye fixations of human test subjects, recorded by eye trackers, is interpreted as ground truth [Riche 13a]. So there exist at least 12 different metrics for comparison between two saliency maps which allow model quality estimation, but many of these metrics correlate between themselves, so according to recommendations given by Riche et al. in [Riche 13a] we need to use at least 3 low correlated metrics. Also, according to Bylinsky et al. in [Bylinskii 16], there are some metrics which we should not omit.

Let us establish explicitly the data to be compared: two saliency maps, identical in size, represented as images in shades of gray: FM – original eye fixation map, produced by eye trackers and considered "ground truth", and SM – estimated saliency map, produced by a model. Then the saliency maps FM and SM are represented by their pixels $FM(x)$ and $SM(x)$ in shades of gray. Let $FM_{\mathrm{I}}(x)$ and $SM_{\mathrm{I}}(x)$ be the gray intensities (in range between 0 and 255) of the pixels respectively. These values represent relative saliency level – the higher is the value, the higher is the predicted possibility of this pixel being salient.

It is also pertinent to mention that similar values of intensity between different saliency maps, produced over different input images, are usually non-comparable due to relative origin of the estimation algorithm.

So let us provide the metrics which we will use in this work, based on the field overview works such as [Riche 13a] and [Bylinskii 16].

First three indicators try to interpret the eye fixation problem as a classification problem, and are based on the concept of receiver-operational curve (ROC) analysis, – graphic interpretation of classification quality, widely used in classification problem.

AUC *(Area under curve) – possibility, that a classifier will more likely correctly classify a random element of "positive class", than incorrectly classify a random element of "negative class".*[2]

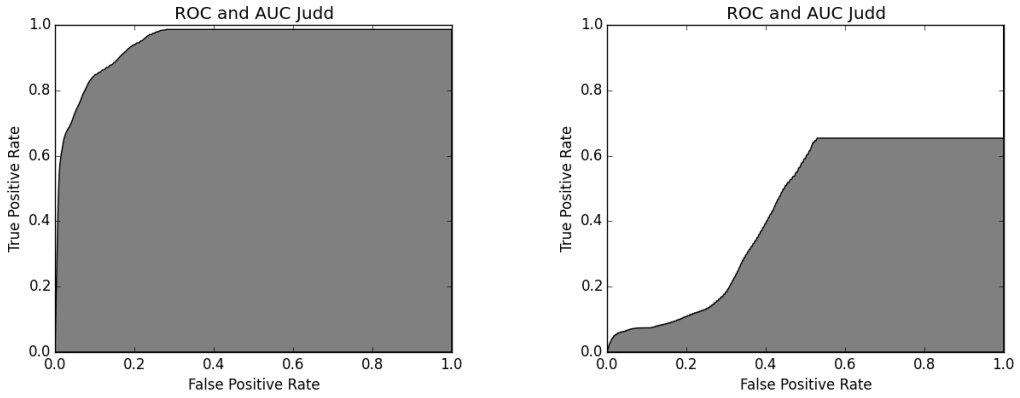

Figure 1.3: Examples of ROC-curve and the corresponding area under curve for an eye fixation saliency map prediction matching perfectly the ground true eye fixation saliency map (left side) and an appalling eye fixation saliency map prediction far from matching the ground true eye fixation saliency map (right-side)

If we treat this definition in terms of eye fixation problem, an element is a pixel of the saliency map; a "positive class" is "pixel is salient", and "negative class" is "pixel

[2]More trivia, definitions and explanation of ROC-analysis in whole and AUC metric in particular can be found in [Fawcett 06]

in non-salient". Such definition is usual for the problems of binary classification:

$$AUC = \int_{-\infty}^{\infty} TPR(T)\, FPR'(T)\, \mathrm{d}T \tag{1.4}$$

Here in Equation 1.4 $\boldsymbol{T}$ is an artificial threshold in mathematical equations of classification model between "positive" (C_0) and "negative" (C_1) classes, which can be manually changed. In [Huang 05] it is shown that $\boldsymbol{AUC}$ is statistically consistent with precision metric, independently of the distributions and bias of classes C_0 and C_1 – which means that high $\boldsymbol{AUC}$ level represents also high precision level, which allows us to use this indicator for the problems, which are not classification problems *per se*.

Most often many problem types are artificially interpreted as a binary classification problem to be able to use an indicator such as $\boldsymbol{AUC}$.

Let us look into these AUC-based metrics, defined for eye fixation problem:

- AUC_{Judd} – AUC-based indicator in interpretation of Judd et al. [Judd 09]. Artificial threshold $\boldsymbol{T}$ is linked to the relative saliency intensity in both images, the "true" saliency map FM and the "estimated" saliency map SM; each possible value of $FM_{\mathrm{I}}(x)$ and $SM_{\mathrm{I}}(x)$ in both saliency maps are taken over a threshold $\boldsymbol{T}$ ($\forall T \in [0; \max_{\forall \mathrm{x} \in \Psi^2}(\mathrm{FM_I(x)}, \mathrm{SM_I(x)})]$, where $x \in \Psi^2$ denotes a pixel and its position in 2-dimensional space), where each pixel $\boldsymbol{x}$ in FM is classified "positive, salient" if $FM_{\mathrm{I}}(x) > T$ and "negative, non-salient" otherwise (same for SM). In this case TPR and FPR for each T value is calculated in general fashion of classification-type problems, checking if "positive salient pixels" from SM correspond in class to the same-positioned pixels in FM, thus yielding true-positive-rates and false-positive-rates for each pair of FM and SM, produced by different values of T.

  A randomly generated saliency map, so called "Chance", would produce $AUC_{Judd} = 0.5$. Values, higher than 0.5, are considered "better, than random". The ideal value is 1.

- AUC_{Borji} – AUC-based indicator in interpretation of Borji et al. [Borji 13c]. Mostly it is almost the same as in AUC_{Judd}, but the difference is that for comparison are taken not all the elements of image (as in general $\boldsymbol{AUC}$), but a random set of fixed size. Also the set should be newly randomized for each new value of $\boldsymbol{T}$, considered for calculation.

- $sAUC$ – "shuffled AUC" in interpretation of Zhang et al. [Zhang 08] is a continued development of the ideas given by Ali Borji while interpreting the

AUC_{Borji} indicator. In the implementation given by [Bylinskii 17] the false positive rate is the proportion of saliency map values above threshold $\boldsymbol{T}$ sampled at random pixels (as many samples as fixations, sampled uniformly from fixations on other images from dataset being considered). As stated previously, usually we cannot consider intersection of different images in comparison; but in this case the accent is given on the average characteristics of dataset in whole – average number of fixation points throughout the dataset, which depends on the quality of eye trackers, number of test subjects, details of the testing protocol, etc.

The third metric, $sAUC$, in several last years became more important, than the first two, in benchmark-based evaluation and validation of models, as the datasets in benchmarks sare thoroughly prepared and protocol-based.

- Another indicator was specifically designed for the problem of comparison of two saliency maps. *The Normalized Scanpath Saliency,* ***NSS*** was introduced to the saliency community as a simple correspondence measure between saliency maps and ground truth, computed as the average normalized saliency at fixated locations [Peters 05]. Unlike in ***AUC***, the absolute saliency values are part of the normalization calculation. ***NSS*** is sensitive to false positives, relative differences in saliency across the image, and general monotonic transformations. However, because the mean saliency value is subtracted during computation, ***NSS*** is invariant to linear transformations like contrast offsets. Considering FMB as a binary map of fixation locations ($FMB(x) = 1$ if $FM_{\mathrm{I}}(x) > T_0$, otherwise $FMB(x) = 0$):

$$N = \sum_x FMB(x) \tag{1.5}$$

$$\overline{SM_{\mathrm{I}}(x)} = \frac{SM_{\mathrm{I}}(x) - \mu(SM)}{\sigma(SM)} \tag{1.6}$$

$$NSS(SM, FMB) = \frac{1}{N} \sum_x \overline{SM_{\mathrm{I}}(x)} \times FMB(x) \tag{1.7}$$

Here T_0 is threshold level, defined by benchmark protocol (sometimes two different sets of ground truth are provided, – saliency map FM and binary map of fixation FMB); $\boldsymbol{N}$ is the total number of fixated pixels in FMB.

A randomly generated saliency map "Chance" produces $NSS = 0$; positive ***NSS*** indicates correspondence between maps above chance, and negative ***NSS*** indicates anti-correspondence.

- Last metric interprets the problem of two maps' comparison as a comparison between two probability distributions. ***KL*** *(K ullback-Leibler divergence)*, also known as *relative entropy* [Contreras-Reyes 12], considers FM as true distribution, and SM as estimated distribution. For calculation purposes we provide normalized version of distributions, NFM and NSM, calculated from values of intensities, interpreted as probabilities. The NSM estimation by NFM is denoted as $NSM||NFM$.

  In this case a divergence $KL(NSM||NFM)$ is defined as a measure of how much information is lost, if we use the "estimation" distribution NFM instead of "real" distribution NSM:

$$NFM = \frac{FM}{\sum_x FM(x)} \tag{1.8}$$

$$NSM = \frac{SM}{\sum_x SM(x)} \tag{1.9}$$

$$KL(NSM||NFM) = \sum_x NFM(x) \log \frac{NFM(x)}{NSM(x)} \tag{1.10}$$

  A randomly generated saliency map "Chance" produces $KL = 2.5$; the higher is the indicator, the worse is estimation. Ideal value $KL = 0$ represents ideal estimation.

Testing eye fixation models have been done using benchmark datasets, such as**Toronto** [Bruce 07], **MIT1003** [Judd 12] and **MIT300** [Bylinskii 17]. Following table depicts best average evaluations of several eye fixation models, both as calculated by us or given in [Bylinskii 17].

In such benchmarks input data is represented as a set of standalone static photos and images, accompanied by ground truth saliency maps and/or binary maps of eye fixation. ***MIT1003*** consists of 1003 testing images (+ 1003 saliency maps and 1003 binary maps), ***MIT300*** consists of 300 validation maps without saliency or binary maps in published.

The biggest problem of models, depicted in Table 1.1 and Table 1.2, is the speed and resource demand. For example, the neural network models demand usage of GPU calculation to be able to provide competitive speed.

The implementations, provided by their creators (mostly implemented in slow interpreted languages such as Matlab and Python), when run on a highly performant consumer computer (Intel i7 3.3 GHz CPU, 32 Gb RAM) *without* GPU calculations, were found pretty slow. E.g., one image of size 1024x768 was processed in a

Table 1.2: Average metrics of the best models, benchmarked over **MIT1003&300** dataset

Algorithm	AUC_{Judd}	AUC_{Borji}	$sAUC$	NSS	KL
DSCLRCN: deep recurrent CNN [Liu 16]	0.87	0.79	0.72	2.35	0.95
SalGAN: generative adversarial networks [Pan 17]	0.86	0.81	0.72	2.04	1.07
BMS: boolean matrix search [Zhang 13]	0.83	0.82	0.65	1.41	0.81
eDN: deep network ensemble [Vig 14]	0.82	0.81	0.62	1.14	1.14
RARE2012: rarity feature analysis[Riche 13b]	0.81	0.80	0.66	1.34	0.89

RARE2012 algorithm was firstly published in 2013, and afterwards improved by Pierre Marighetto at LSUN SALICON challenge in October 2015

time ranged from 2 seconds (Cheng's global contrast R-model) up to 200 seconds (SalGAN).

While we say that these models are high-demanding in terms of hardware and implementation, they are still very efficient. Given time, as embedded hardware will gain computational abilities, these models will be able to be implemented directly in autonomous mobile platform; yet, nowadays we need to relay onto classical approaches, which perform not really worse, but can be run onto modern mobile platforms.

1.4 Object Recognition and Semantics

The prevailing view, according to [Borji 13c], is that bottom-up and top-down attentions could be combined to direct the attentional behaviour. An integration method should be able to explain when and how to attend to a top-down visual item or skip it for the sake of a bottom-up salient cue, which implies the importance of some kind of a decision mechanism, based on different possible cognitive conditions.

A decision mechanism can be based on any set of parameters and/or techniques; one of the approaches here can be an ensemble of lexical labelling and semantic analysis – *what exactly do we see here and is it really important for us now.*

For an autonomous mechanism, a problem of such labelling self-reduces to the problem of visual object recognition – itself a big problem with a corpus of existing research on the topic. While this problem is not yet really solved, it is definitely out of the scope of this work; we can make only a brief overview of existing approach

classification, and provide some justification for usage of some approaches over others based on the conditions of the context of this work.

Same goes for the problem of semantic analysis, – also pretty popular in the field of artificial intelligence. For a decision mechanism in a model of an intrinsic eye mechanism we do not need a full-pledged analysis. A simple superficial mechanism could suffice in order to find the usability of either the top-down or bottom-up result for each case.

1.4.1 Existing Approaches in Object Recognition

While the word "recognition" both in human neurology and computer vision is taken as granted, it provides several meanings, where the most used is a *"providing a label for some visual scene or object"*; although, there does not exist a full classification neither of the meanings nor of the usages. Mostly the researchers ask themselves *"how to recognize something"*, or *"what part of human brain is involved in recognition"*.

But there could be other questions, such as, for example, *"What can a human visually recognize?"* Such a simple question could be answered in several points [Walker 02]:

- *Something, that has been recently learned,* – and recognized specifically, as as an exact object (or an object, similar to it), – e.g., just learnt letter can be easily recognized by a child, or a freshly viewed painting. In terms of computer vision, – an object, seen previously and "memorized", should be able to be recognized;

- *Something, that is similar to some well-known patterns,* – such as human face, human body, tiger stripes, etc. This type of recognition is not only high-level cognitive, but also very basic and instinctive. According to [Borji 13c], the objects with additional information "embedded" into them, would be almost always more salient, than other objects. Such visual cues with additional info "embedded", for example, are written words and symbols (which are not only "lines and scribbles", but also have additional semantics), or human faces (which hold additional information of, e.g., emotions, or a possibility of interaction);

- *Something, that is similar to some objects previously learnt,* – and this type of recognition provides more broad set of categories and labels (note the difference of definitions between first and third points, as "recently" is not equal to "previously", and neurologically involves different parts of memory, short-term

as opposed to long-term). In terms of computer vision, this type of recognition is usually thought of as a default type, as it is possible to implement the long-term memory as a dataset, or database, or any abstract information-holding structure available in informatics.

1.4.1.1 Keypoint-based recognition

Starting from initial article [Lowe 99], describing the "Scale-invariant feature transform" (SIFT), there emerged a family of algorithms, which are based on keypoint detection, description, and descriptor comparison, which found its applications widely in computer science.

According to the comparison by [Miksik 12], the list of algorithms could be narrowed to 4 best performing algorithms:

- "Scale-invariant feature transform" (SIFT) [Lowe 99] – keypoint detector, based on difference of Gaussians, with decimal value descriptors;
- "Speeded-up robust features" (SURF) [Bay 08] – SIFT modification, which performs on a comparable level, but provides faster computation;
- "Oriented fast and Rotated BRIEF" (ORB) [Rublee 11] – a modification of BRIEF algorithm, which introduced the idea of binary descriptors in order to simplify the calculations, – as in this case for a comparison it is sufficient to calculate Hamming distance, and not Euclidean;
- "Binary Robust Invariant Scalable Keypoints" (BRISK) [Leutenegger 11] – another binary descriptor-based algorithm, declared as the most efficient in the family by its authors.

We provide the comparison of these four algorithms over two different datasets, involving an interpretation of Precision and Recall, as shown in Table 1.3.

In this table the indicators are given in form of Precision and Recall in Christian Wolf interpretation [Wolf 06], introduced and widely used in ICDAR competitions [Lucas 03].

The datasets Graffiti and Pascal VOC (combined into one dataset **Graffiti/PVOC**) are used in the aforementioned comparative studies, where the evaluation metrics (second part) are taken from. Another dataset, **Things-50**, has been produced by us (more details in Appendix A), and so are the evaluation results.

According to Table 1.3, leading with a slight advance, the **BRISK** algorithm have shown better efficiency, which gives us a justification to use this model as a

Table 1.3: Comparison of four keypoint-based recognition algorithms over two datasets, Things-50 and Graffiti/PVOC

Algorithm	Dataset	PR_{Wolf}	REC_{Wolf}
SIFT	Things-50	0.491	0.463
SURF		0.543	0.581
ORB		0.494	0.474
BRISK		**0.546**	**0.592**
SIFT	Graffiti/PVOC	N/A	N/A
SURF		0.485	0.513
ORB		0.493	0.495
BRISK		**0.504**	**0.527**

keypoint-based technique, usable in "fine-grain" recognition element of recognition module.

1.4.1.2 Pattern-based recognition

Since initial works in the field of visual saliency in social robotics (e.g., [Breazeal 99]) several patterns have always been tamed as "pop-out" social objects which always attract more attention. One class of such "pop-out" objects are human faces; another can be named as "texts" – an ordered set of symbols of a known alphabet which can be treated as words with meanings.

Problem of detection of faces or texts in real time also has a set of robust algorithms to apply, as Viola-Jones framework of Haar-like features' cascades for detection of distinctive patterns inside a shape (as a human face), or as Chen's algorithm of MSER and Canny edges for detection of texts.

While for face detection the Viola-Jones framework [Viola 04] since its inception in 2004 has become an industry standard de facto due to its robustness and efficiency, the real research in the field has undergone a massive "desolation" – there exist only around ten top-cited articles, dated after 2004, with keywords "face detection" in the name, and mostly they are concerned on other different contexts of this task. Due to such a "desolation" in the field, there is not much to compare in terms of a task of face detection.

On the other hand, the problem of text detection and recognition in different contexts has been flourishing, always present in ICDAR (International Conference on Document Analysis and Recognition) competitions with up to 100 (and even more) new models submitted each two years.

E.g., only at ICDAR 2015 competitions [Karatzas 15] there firstly arose the challenge of localisation of "incidental scene text" – text, seen in real world images,

without any prior knowledge or preparation of data.

But most of the submitted models neither have open implementation, nor even have an article about it, due to the models' authors preferring a commercial possibility of the models' usage — thus, in ICDAR competition resulting table, if sorted by integral metric of model quality, the best model with open implementation (as seen in GitHub), *"EAST: An Efficient and Accurate Scene Text Detector"* [Zhou 17] is only on the 19-th place, while still performing very well ($PR_{Wolf} = 77.32\%$, $REC_{Wolf} = 84.66\%$), and among more than 70 models there are less than 10 models with open implementation and article. Several examples of image processing by the EAST algorithm are given in Figure 1.4, where cyan rectangles depict the detected text regions.

Figure 1.4: Cropped examples of EAST algorithm results on the V-60 dataset

In general, quasi-total number of these models are based on convolution neural network approach, mostly treating the text detection task as object detection.

Concluding thus subsubsection, we can state that usage of both Viola-Jones framework and EAST CNN for face and text detection respectively brings us the pattern-based tier of this recognition module.

1.4.1.3 Broad category recognition

Another category for object recognition – "broad recognition", – is the recognition of unknown objects, based on previous learning. This field has also shown in several previous years a big leap in the efficiency and robustness of the approaches. The absolute leader de-facto is, again, a family of algorithms, based on convolutional neural networks due to its robustness against input in form of images and

its deep potential of generalization and distinguishing several thousands of different categories.

This question here is a main type of problems on annual academic and research challenge ILSVRC (ImageNet Large Scale Visual Recognition Competition) [Russakovsky 15], – Similar to ICDAR competitions in the field of text detection and recognition, – a competition, where several hundred thousands images are categorized into 1000 arbitrary categories, based on the depicted object, in terms of previously mentioned WordNet/ImageNet synsets. This means that we can apply one or several winning or runner-up models from this challenge and be able to change the model as new winning researchers emerge through time.

Although each year the competition changes the main challenge, the ILSVRC2012 dataset is considered as "standard" dataset in the field due to the choice of categories, represented in the sample, and its high level of distribution over WordNet lexical tree. The ILSVRC2012 dataset has been considered as training and validation dataset not only in 2012, but in the following competitions as well along with other new datasets.

To be able to choose one of the dozens existing models, submitted to ILSVRC competition, we can create a "short list" of existing winners and top runners of the competition, – in contradistinction with ICDAR competition, most of the winning models are open in both implementation and theory.

Then we can make comparison by ourselves, both on existing **ILSVRC2012-Val** (validation) dataset, as well as our **Things-50** dataset.

In 2014 competition on ILSVRC2012 sample dataset were found, as winners, three models: GoogLeNet [Szegedy 15], VGG [Simonyan 14] and MSRA (also known as ResNet [He 14]). It is also pertinent to include the well known top-runners NiN [Lin 13], AlexNet [Krizhevsky 12] and BVLC-AlexNet modification (provided by the creators of well-known open source robust platform for deep neural networks, CAFFE [Jia 14]).

Let us define the experiment protocol: the algorithm processes several images, and for each of the images it produces 5 recognition categories with highest confidence level CL ("TOP-5" recognition), and also chooses 1 category out of these 5 ("TOP-1" recognition). If, among "TOP-5" categories, one is correct category, this image is interpreted as correctly classified in terms of "TOP-5" challenge, else – as classified incorrectly. Same goes for "TOP-1" challenge.

Assuming, that there exists K images, each representing one object with only one (out of 1000) correct classification category, the main classification metric for j-th model, recall $REC(j)$, can be defined as shown in Equation 1.11, where $CC(i, j)$ represents a binary function: 1, if i-th image is correctly classified by j-th model, 0

otherwise.

$$REC(j) = \frac{\sum_{i=1}^{K} CC(i,j)}{K} \tag{1.11}$$

Such metric works for both, "TOP-1" and "TOP-5", challenges, and will surely be less for the first challenge, than for the second.

Results of our experiments are shown in Table 1.4. While Microsoft neural network ResNet-152 shows almost fantastic efficiency on the ImageNet validation dataset, it proves itself pretty helpless, along with all the other models, on a pretty specific dataset Things-50, which was aimed towards exactly recognition in terms of ILSVRC2012 categories, but with a bias towards the household/office objects in both controlled (neutral) and chaotic environments.

It is also pertinent to note, that while **ILSVRC2012-Val** contains 50000 images, pretty fairly distributed into 1000 categories, the statistical bias and small size of the sample of **Things-50** doesn't provide an opportunity to affirm that ResNet is better or worse than VGG-19 in household or office environment; yet, it shows that such a possibility is plausible and we have to consider a change of models, if the overall task, or at least some conditions of the task of the visual attention model changes.

1.4.2 Language Analysis

As already stated previously, the way of labelling, used in the broadest recognition problem – the convolutional neural networks, taught upon ImageNet image dataset (as stated previously in subsubsection 1.4.1.3), is the universal word classification.

The ImageNet is organized as billions of images, each depicting a concept or an object, and labelled as such. The labels are categorized thoroughly into a giant semantic tree, taken from WordNet [Miller 95] (already mentioned in Table 1.3 as the provider of "synsets" and their codes).

In short, **WordNet** is a lexical database for the English language. It groups English words into sets of synonyms (called synsets), provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. This ontology gives us an opportunity in language analysis, as WordNet applies different types of relations onto nouns, including:

- **hypernyms:** Y is a hypernym of X, if every X is a (kind of) Y (*canine* is a hypernym of *dog*);

Table 1.4: Broad category recognition models comparison over **ILSVRC2012-Val** and **Things-50** datasets

Model	Dataset	REC_{TOP1}	REC_{TOP5}
VGG-19[1]		0.6878	0.8803
VGG-CNN-S[1]		0.4824	0.7241
ResNet-152[3]		**0.755**	**0.9167**
GoogLeNet	ILSVRC2012-Val[2]	0.5225	0.7699
BVLC-AlexNet		0.3678	0.6129
AlexNet		0.3638	0.6073
NiN		0.4181	0.6718
VGG-19		**0.4706**	**0.6275**
VGG-CNN-S		0.3333	0.51
ResNet-152		0.451	0.5686
GoogLeNet	Things-50	0.3725	0.549
BVLC-AlexNet		0.255	0.51
AlexNet		0.2157	0.49
NiN		0.1765	0.2941

[1] VGG-19 and VGG-CNN-S are considered as different "flavours" of general VGG model and should be assessed independently

[2] **ILSVRC2012-Val** dataset is provided by ILSVRC organisers specifically for model validation

[3] ResNet-152 is the most recent modification of MSRA/ResNet network, and considered by its authors as the best in the family

- **hyponyms:** Y is a hyponym of X, if every Y is a (kind of) X (*dog* is a hyponym of *canine*);
- **coordinate terms:** Y is a coordinate term of X, if X and Y share a hypernym (*wolf* is a coordinate term of *dog* and vice versa);
- **meronym:** Y is a meronym of X, if Y is a part of X (*window* is a meronym of *building*);
- **holonym:** Y is a holonym of X, if X is a part of Y (*building* is a holonym of *window*)

Semantic similarity based on WordNet has been widely explored in Natural Language Processing and Information Retrieval. But most of these methods are applied in an ontology (e.g., WordNet).

Several methods for calculating semantic similarity between words in WordNet exist and can be classified into three categories:

- **Edge-based methods:** to measure the semantic similarity between two words is to measure the distance (the path linking) of the words and the position of the word in the taxonomy. That means the shorter the path from one node to another, the more similar they are (e.g., Wu-Palmer distance [Wu 94] in modification of Pedersen [Pedersen 04]).
- **Information-based statistics methods:** to solve the difficult problem to find a uniform link distance in edge-based methods, Resnik proposes an information-based statistic method [Resnik 99]. The basic idea is that the more information two concepts have in common, the more similar they are. This approach (and its relatives) is independent of the WordNet corpus and demands additional information sources.
- **Hybrid methods:** combine the above methods, using both the "tree path" between concepts in a tree of relations, and information content, e.g., [Jiang 97].

While all of the similarity metrics address ontological similarity, they do it in different way. Thus, only three of them represent a normalized metric which produces a value in range [0..1], so if we would like to use such a metric as a quantitative modifier, we need to choose among them all.

Wu-Palmer similarity is defined as the similarity of two concepts based on the common concepts by using the path, as shown in Equation 1.12. Here C_3 is the least common superconcept of C_1 and C_2. N_1 is the number of nodes on the path from C_1 to C_3, N_2 is the number of nodes on the path from C_2 to C_3; N_3 is the number of nodes on the path from C_3 to root.

$$WUP(C_1, C_2) = \frac{2 * N_3}{N_1 + N_2} \tag{1.12}$$

Lin similarity is a score denoting how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node) and that of the two input concepts. The relationship is given by the Equation 1.13, where $IC(C_n)$ represents function of information content: the conventional way of measuring the IC of word senses is to combine knowledge of their hierarchical structure from an ontology like WordNet with statistics on their actual usage in text as derived from a large corpus (e.g., Brown corpus [Francis 79]).

$$LIN(C_1, C_2) = \frac{2 * IC(LCS)}{IC(C_1) + IC(C_2)} \tag{1.13}$$

Path similarity represents a very simple metric of how long is the path in the ontologic tree, as shown in Equation 1.14, where $PL(C_1, C_2)$ represents path length between concepts in ontologic tree with edges representing hypernym/hyponym relations.

$$Path(C_1, C_2) = \frac{1}{PL(C_1, C_2)} \tag{1.14}$$

An example of similarity calculation between concepts "dog" and "wolf", along with "dog" and "cat", is shown in Listing 1.1.

```
T1 = HyperTrees( dog ) =
    [1] {*ROOT* < entity < {8 concepts omitted} < mammal <
    placental < carnivore < canine < dog}

T2 = HyperTrees( wolf ) =
    [1] {*ROOT* < entity < {8 concepts omitted} < mammal <
     placental < carnivore < canine < wolf}

T3 = HyperTrees( cat ) =
    [1] {*ROOT* < entity < {8 concepts omitted} < mammal <
     placental < carnivore < feline < cat}

LCS(T1, T2) = { canine }, depth = 14
LCS(T1, T3) = { carnivore }, depth = 14

DepthT1 = min(depth( {tree in T1 } )) = 15
DepthT2 = min(depth( {tree in T2 } )) = 15
DepthT3 = min(depth( {tree in T2 } )) = 15

wup( dog , wolf ) = 0.9333
wup( dog , cat ) = 0.8666

IC( carnivore ) = 7.2549003421277245
IC( canine ) = 7.638625463599483
IC( dog ) = 7.7404081579094255
IC( cat ) = 8.630265632715425
IC( wolf ) = 11.072612668084629

lin( dog , wolf ) = 0.8121
```

```
lin( dog , cat ) = 0.8863

Shortest path: {
    "subsumer" : "canine",
    "lpath" : "dog",
    "rpath" : "wolf" }
Path length = 3

Shortest path: {
    "subsumer" : "carnivore",
    "lpath" : "canine < dog",
    "rpath" : "feline < cat" }
Path length = 5

path( dog , wolf ) = 0.333
path( dog , cat ) = 0.2
```

Listing 1.1: Examples of different similarities between close concepts

As it can be seen in Listing 1.1, Lin similarity is dependent onto additional metric of "information content", thus making similarity between "dog" and "cat" higher, than between biological relatives "dog" and "wolf". Such a non-obvious approach to calculation can jeopardize autonomous decisions of a robot, – while being pretty interesting and emergent of a non-linear approach, such a behaviour is beyond the scope of this work.

On the other hand, the Path similarity shows hyperbolic diminution, and for even very similar concepts gives quantitative value lower, than 0.5; while the difference between really far concepts (such as noun "cat" and noun "love" in their first meanings) path length is 20, thus scoring $path(cat, love) = 0.05$.

```
T1 = HyperTrees( cat ) =
    [1] *ROOT* < entity < physical_entity <
    {7 concepts omitted} < mammal < placental <
    carnivore < feline < cat
T2 = HyperTrees( love ) =
    [1] *ROOT* < entity < abstraction < attribute < state <
     feeling < emotion < love
T3 = HyperTrees( glove ) =
    [1] *ROOT* < entity < physical_entity <
    {4 concepts omitted} < clothing <
```

```
    handwear < glove

wup( cat , love ) = 0.1739
lin( cat , love ) = 0.0
path( cat , love ) = 0.05

wup( cat , glove ) = 0.4
lin( cat , glove ) = 0.0
path( cat , glove ) = 0.0625

wup( glove , love ) = 0.2222
lin( glove , love ) = 0.0
path( glove , love ) = 0.0667
```

Listing 1.2: Similarity between far concepts

As it is depicted in Listing 1.2, Lin similarity drop to 0 when the LCS is a very common concept; while Path score shows insignificant results of a 10^{-2} magnitude, depicting a bias towards low-level scores.

Wu-Palmer similarity, in contradistinction to others, produces significant evaluation both in cases of close concepts and far concepts, giving a big gap between these estimations, and being pretty monotonic with common sense. Thus, according to Wu-Palmer estimation, a comparison of 5 concept pairs ("dog-wolf" > "dog-cat" > "cat-glove" > "cat-love" > "love-glove") shows an ordering by generalization of common ancestor. Given a weak hypothesis, that this ordering correlates with "common sense", we can apply a survey method to partly justify it.

We have done a small survey among 40 students. They were given these 5 pairs of concepts and asked *"to place them in order of decrease of similarity"*. The results show that more than a half (68% at least) agree with such an ordering, implying it to be a quasi-"common sense"; brief results of this survey are depicted in Figure 1.5. While the question of "common sense" is mostly rhetorical on one hand, and philosophically very large on the other, the bigger scaled survey method could approach it: the bigger set of concept pairs, along with the bigger pool of interviewees can justify more strict hypotheses; yet, such question is not in scope of this work.

The given survey can (at least partly) justify usage of Wu-Palmer metric as an answer to a question – *"If the task implies a search of **A**, how good it is to find **B** instead?"*.

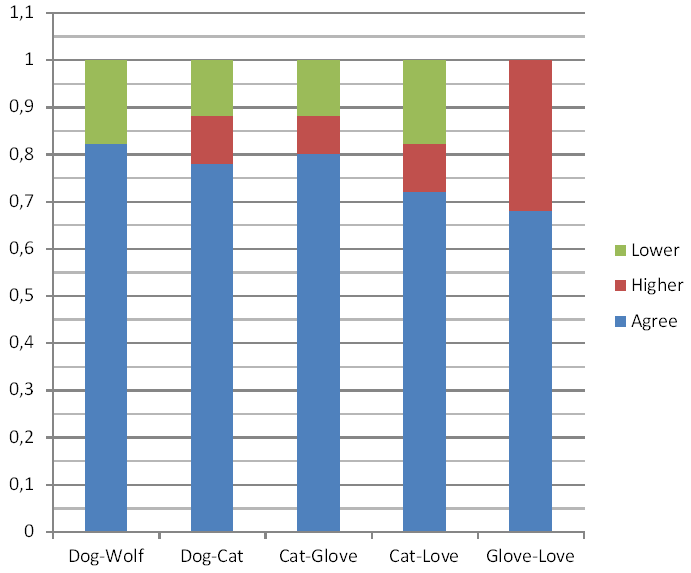

Figure 1.5: Results of a mini-survey, whether the given order of concept pairs (by decreasing of Wu-Palmer similarity) is in common sense; "Higher" means that the interviewee placed the pair higher in position, "Lower" – that interviewee placed the pair lower in position

1.4.3 Knowledge storage

Another implicit part of decision making module is some storage for existing knowledge, as we need to categorize the data already found, and to differentiate it from the data newly acquired.

The models of memory or knowledge storage can be informally divided into two classes [Gavrilova 00] – the "hard models", which could be defined by a formal language, and "soft computing models", as defined by [Zadeh 94], which are applied for inexact solution of computationally hard problems.

While the "soft computing" approach is directed more towards data processing, the neural networks provide several types of memory-wise networks, including Hopfield network [Hopfield 82], bidirectional associative memory (BAM) [Kosko 88], Boltzmann machine [Ackley 85] or previously mentioned convolutional neural networks [Matsugu 03].

As showh in [Hertz 91], the network capacity of the Hopfield network model is determined by neuron amounts and connections within a given network. Therefore, the number of memories that are able to be stored is dependent on neurons and connections, and the recall accuracy between vectors and nodes was 0.138 (approximately 138 vectors can be recalled from storage for every 1000 nodes) [Liou 06]. Similar constraints apply to BAM and Boltzmann machine.

Soft computing approach is not very suitable for the task of short-term memory

of exact objects – along with the characteristics of these objects and possible meta-data; such a task call more for a "hard model". Throughout the formal models, there have been three meta-techniques, usable for memory or knowledge storage:

- Production rule systems [Brownston 85];
- Semantic networks [Scragg 76];
- Frame networks [Minsky 75].

Production rule systems. A defining element of a production rule system is representing knowledge in form of a set of rules *"If **A**, then **B**"*, which allows, according to a set of input data, according to the constraints, apply actions – e.g., induce new data or knowledge, preconize an expert verdict, etc. Main advantage of such a model is flexibility and simplicity of the output mechanism, along with being "sharpened" exactly for actions.

Such models are popular in industrial expert systems [Klahr 87, Khoroshevsky 93], along with the tasks of teleo-reactive control of agents in multi-agent systems [Nilsson 94, Hayes 08]. Such models are usable in context of the tasks of autonomous robotics, but more in the specificity of actions and reactions. Such models demand "pre-learnt" set of knowledge (usually created by a human expert). In this case it is expedient to apply such models in terms of semi-free scenarios, as a set of actions-reactions for specific events or specific visual images, or estimation of expert verdicts according to existing requests. Most known example of such application in AI is IBM Watson, which solves problems in several contexts according to very complex production-rule-based decision systems.

Semantic networks. Semantic nets, or networks, are structures which could be represented by an oriented graph, where its vertices are concepts, and its edges are relations between concepts. As a **concept**, one could define abstract or specific objects; as a **relation**, one could define any linkage between concepts. Usually, default relations in semantic network systems are "AKO" (a kind of), "Has part", etc.

Such structures, differently of the production rule systems, are not oriented onto solution creation, – more onto description of existing situation in given environment. In this case, a typical "semantic network problem" is defined as a search of such a fragment of the network, which corresponds to some input conditions. Main advantage of such knowledge organisation is that it is more, than others, suited according to contemporary representation of both short-term and long-term memory organisation [Scragg 76].

The main disadvantage in this case is the difficulty of search process.

Most often the semantic nets are used for storage of knowledge about objects, than for induction of new knowledge. However, they are also used in different expert systems as knowledge presentation language [Durkin 93].

Frame networks. Frames are defined as abstract images, used for representation of a perception stereotype. Such images are averaged stereotypes of defined objects; thus, abstract image "child" is an aggregate of characteristics, inherent for each child, without any further detail (*"child is a human, which is less than 18 years old"*). Each frame can be then called as a set of properties, inherent to this image.

There exist "sample frames", depicting an abstract concept, and "instance frames", describing additional details of each independent object, which is an instance of existing abstract image, and inherits its properties.

Frames apply a classification upon many properties; also, they are linked via AKO-relations, which makes it possible to inherit the properties of a more abstract frame. Thus, frames are more complex structures, partly resembling semantic nets, which describe the world more fully and flexibly [Sowa 14].

As the model quality assessment in the field of knowledge representation exists only on the level of estimation, if the induced solutions of expert systems are wrong or right [Gavrilova 00], the choice of one or another model to be used in solving each independent task should be done on the basis of different premises.

1.5 Conclusion

This chapter has given the reader an overview of the field of visual attention, including best techniques, used in the domain of visual attention, and best techniques in adjacent domains, such as object recognition and simple lexical comparison. We provide also the means of evaluation and do comparison for the state-of-art approaches efficiency, as applied to the synthetic benchmarking datasets.

The best techniques in the field, mostly based on deep neural networks, have caveats if applied to autonomous robotics – the inability to run in real time due to high resource demand. This leads us to the choice of classical bottom-up approach based on the compromise "resource supply / quality of solution" as a key part of the visual attention system.

Another key point of the chapter, where we need to emphasize the reader's attention, is a giant gap between two attention directions, the bottom-up and top-down, which are derived from absolutely different reasonings: elementary vision cues against high-level cognitive phenomena. There have been a handful of approaches

towards crossing this gap, trying to integrate both directions into one combined approach – e.g., [Navalpakkam 05], or [Peters 07]. Most of them implement the top-down direction as a "black box" using methods, such as Fourier transformations, which are not aware of cognitive processing.

While one could argue about the usage of non-cognitive methods in this matter, it should be noted that the state of modern hardware and algorithmics has finally approached the moment where a top-down attention can be implemented as a complex model of artificial attention with some partial modeling of human brain's lateral prefrontal cortex in the context of computer and machine vision [Buschman 07]. Which is why we can aim our research in this direction, – combination of two attention directions, where top-down model is more sensible and cognitive-like.

The means for such a cognitive-like approach could be provided by aforementioned best techniques from adjacent fields. Quasi-classification algorithms in the field of object recognition (which can be reformulated as lexical labelling) and simple lexical comparison for similar or relative concepts, can be used in order to provide some distinction and a basis for a simple decision module, essential for any combination of top-down and bottom-up attention models.

It is in this direction that my research is aimed; Chapter 2 briefly outlines the general scheme of such a combined model; Chapter 3 details this system both on top-down and bottom-up parts; Chapter 4 is dedicated to the implementation, validation and several real-world tasks.

The next chapter considers the general outline of such a system, representing its theoretical framework, in order to achieve the general goal, as shown in General Introduction.

2 | General Outline for Combined Visual Attention Model

2.1 Introduction

As discussed previously in subsection 1.2.1, visual attention models are usually divided in two big groups: ”bottom-up” stimulus driven models, which usually try to solve a task of free visual search, and ”top-down” expectation driven models, which are influenced by cognitive phenomena (such as knowledge, task, reward, etc.). The Chapter 1 concluded on a note that a combined visual attention model could be created.

If we would try to create a ”generalized” implementation for autonomous or semi-autonomous usage in robotics, we need to cover as most conditions as possible: not always a ”bottom-up” attention model will be usable (e.g., when a robot has a task to navigate to a human), and not always a ”top-down” attention model can help (e.g. when a robot is in a semi-empty scene, surrounded only by several unknown objects, and doesn't have any particular task).

The question of combination of ”bottom-up” and ”top-down” models has been explored for several years already, but each time in limited context of research and not totally coherent with the scope of this research – e.g., reward-driven choice making [Navalpakkam 10] or video game eye fixation prediction [Peters 07].

This chapter is directed towards outlining a possible system, combining both approaches, as well as our guided choice for the techniques and approaches in the parts of such system, – if such choice is possible.

A ”bottom-up” model is usually computationally more simple, than “top-down”, thus it can be embedded into a modern robot; yet, as any complex ”top-down” approach needs computational strength, we can approach it in a ”weaker” sense, choosing remote architecture for this part of system. This gives us an opportunity to use the most of existing approaches, which solve this or that part of the big problem; yet, we have to find or define the means to evaluate the approaches and techniques, used in this model construction, in order to be able to choose between similar ana-

logues, so we also define several datasets, used in this paper, and evaluation metrics, used above them.

So, the ground rules of the system implementation can be lay like this: while its basic, "bottom-up" part still has a constraint to be able to run on robot's CPU in quasi-real-time, the second, "top-down" part, can be implemented in terms of "cloud consciousness", where robot asks for "help" via network to the applications, which are physically situated elsewhere. Such approach would allow us to still have a quasi-real-time usage experience, combined with best techniques in image processing (usually based on convolutional neural networks, which demand GPU-based processing).

The section 2.2 outlines "grand scheme" of a Combined Model, along with several techniques which could be used in it. Providing description and some general consideration about usability, are section 2.3 for the bottom-up unit, section 2.4 for top-down unit, and section 2.5 for a decision module, in order to establish the outline of the system and its units. The chapter is concluded with outlines of possible extensions and validation directions of the model.

2.2 An Outline of a Combined Visual Attention Model

As we have already discussed in Chapter 1, a combined visual attention model would consist of three main parts: top-down unit (TD), bottom-up unit (BU), and a decision making unit, which should use a set of weak conditions to choose between TD and BU results. In Figure 2.1 it is shown such a structure.

As we stated in the last paragraph of subsection 1.2.2, in order to achieve a possible autonomy in modern robot, we need to implement the BU part as computable on modest resources of embedded CPU. Thus we have to rely on classic, contrast-based eye fixation approach, as opposed to currently state-of-art convolutional network-based approaches.

The TD part should be generally task-driven due to the context of the work, which means the necessity of object detection and recognition on several levels.

2.3 Bottom-Up Unit

An overview of the human attention modelling problem was made in [Borji 13c], establishing a classification of subtasks in this field and differentiating more than 60 different models into 28 subsets of methods. As previously discussed in subsec-

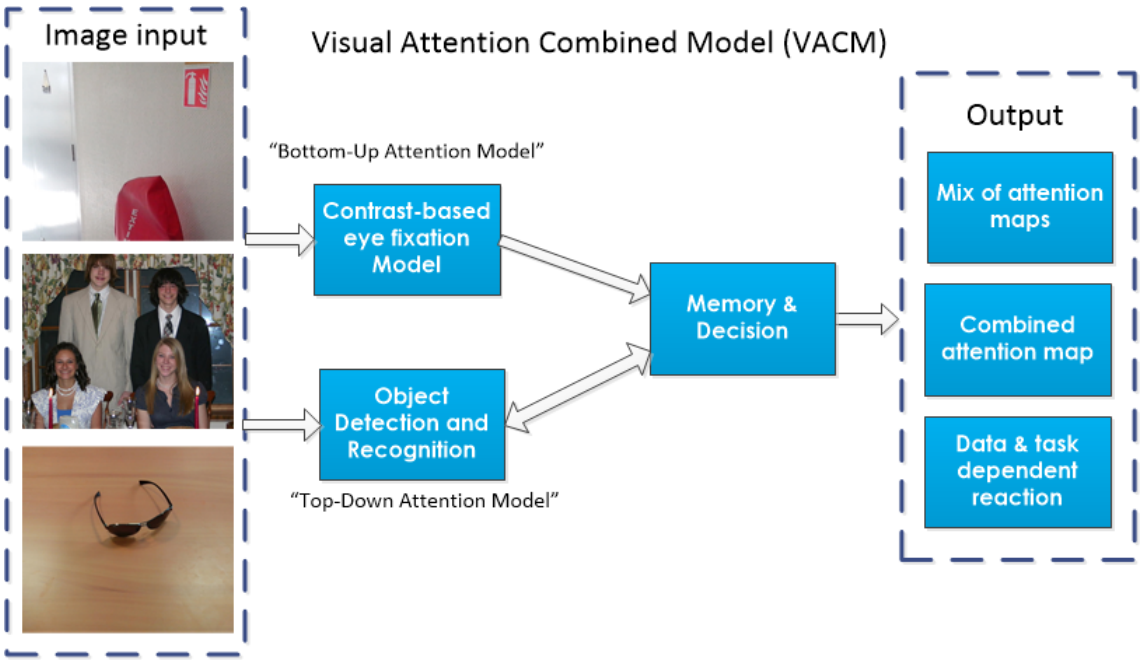

Figure 2.1: An extended human-like visual attention model, combining both attention directions

tion 1.2.1, some of these subsets depend (at least partially) on detection of salient objects in scene [Liu 11, Navalpakkam 06]; other subsets deal with detecting (or defining) the most interesting spot in scene, which may or may not be a part of the most salient object of this scene [Kadir 01, Kienzle 09].

And a number of listed models covers different subtasks. For example, the approach described in [Zhang 13] covers both object detection problem and eye fixation prediction problems.

In the same way, a model which is targeted onto a specific task – e.g., fire detection, – or more generally the detection of complex, deformed and multi-shaped objects, may cover "scene classification", "salient object detection" or "interesting point detection" (accordingly to classification of subtasks presented by Borji), while, according to [Võ 12], not being any of the above-mentioned tasks, but representing somehow some similarity with the task of eye fixation prediction due to high overall saliency of the fire, which, with high probability, will be an eye-fixation point in images containing fire.

The idea is based on establishing some similarity between the salient vision and the human's natural vision (namely through human's eye fixation problem set paradigm), then exploiting the ascertained likeness in order to make machine vision come closer to human-like vision. In fact, although focusing distinct purpose and different applications, the eye fixation paradigm provides appealing source linking

human-like visual skills.

2.3.1 Analogy between Salient Object Detection and Eye Fixation Tasks

In order to defend the hypothesis of interusability of the algorithms in two similar and related, yet still different contexts of problems, – salient object detection and eye fixation, – we take the algorithm related to the previous problem, solved in the LISSI laboratory: statistic-driven center-surround concept for objects detection, as defined in [Ramík 12] and largely inspired from introduced notions. As that algorithm could not handle the aforementioned purpose of visual attention modelling, we will have to take only the first-step part of it while checking its ability to solve given tasks in new context. As in fact, in the above-mentioned objects detection concept, even though constituting a leading-step, the so-called saliency detection acts as intermediary process in order to extract (e.g. detect and isolate) the salient objects in an image. In other words, the primary action and thus the main outcome of the old algorithm in this case was to detect and to isolate the potential salient items without any linkage to human's way of focusing objects in a landscape or to the human-like visual attention.

We have considered the eye fixation problem paradigm as evaluation benchmark for experimental validation of investigated approach. As the goal in the eye fixation problem paradigm is modelling the human eye fixation (supposed linking human visual attention) when watching a scene (namely an image), the main applicative goal is the prediction of humans' center of attention versus the presented image. Most applications deal with either the accurate design of web pages' visual content or relate the design of visual content for advertisement issues (commercial applications). Although focusing distinct purpose and different applications in this chapter, the considered paradigm provides appealing experimental credentials linking "human-like" visual skills. In fact, stating that the human eye-fixation mechanism tends following what may be considered by the human as being relevant in visually perceived information, one may establish (or state on) some similarity between the robot's artificial vision and human's natural vision.

However, in order to establish the comparison criteria (indicators and metrics) within the considered experimental frame, we need to make a two-way appraisal: first treating the eye fixation paradigm as an object detection problem, and then treating the object detection algorithm as eye fixation predictor.

2.3.1.1 Eye Fixation Problem as Quasi-object Detection Problem

Let's consider the eye fixation map as an image that pixels' intensities represent the eye fixation probability (interpretable as being some kind of pixels' visual-attractiveness-degree) – also defined previously as saliency map in the eye fixation problem paradigm. Assimilating the eye fixation map to such an image, one can define a quasi-object as an item, represented by a set of contiguous pixels (of the image) highlighted by relative probabilities (or intensities) higher than some threshold T. In other words, a map of such quasi-objects can be obtained by applying an arbitrary threshold to an eye-fixation map. On the other hand and in the same way, a quasi-object of such an image could also be seen, from the saliency detection point of view, as a salient object (or salient region) versus the rest of the image considered as background.

Taking into account the latest analogy, the adaptive-threshold-based salient objects' detection algorithm could be used to detect salient quasi-objects in MIT1003 database images. Figure 2.2 shows examples of quasi-objects carried out using the aforementioned algorithm for two different threshold values ($T = 10$ and $T = 50$) as well as the eye fixation map obtained from experimental setup involving various groups of humans watching the concerned stimulus (input image).

In this case we can compare the detected areas to salient regions, obtained from MIT1003 fixation maps. For regions detection we apply the GSM/LSM/FSM algorithm with default parameters $\{IRC = 0.2; WSC = 0.4\}$. As an evaluation we can apply usual Precision and Recall technique, as in pure object detection problems, due to similarity with classification problem [Riche 13a].

In Figure 2.3 are shown Precision and Recall values for different thresholds.

By comparing the so-called detected quasi-objects to the most visually-attractive zones in the corresponding eye fixation map (characterized by precincts representing high probabilities of eye fixation), it is pertinent to note the analogy between the detected salient zones and the most relevant eye fixation precincts.

The Recall value in this case can be interpreted as "ratio of precincts focusing effective human attention through his frequent eye fixation on these regions detectable by object detection algorithm as salient regions". In Figure 2.3 are shown Precision/Recall values versus different thresholds over whole 1003 images of MIT 1003 eye fixation benchmark database. It is pertinent to note that 80% (and at least 68%) of detected quasi-objects (salient zones) match with the eye fixation precincts and thus, are detectable by the algorithm. It also highlights that at least 20% of detected salient regions do not represent eye fixations over the detectable salient objects. The reported results show also that a suitable tuning of the threshold, and

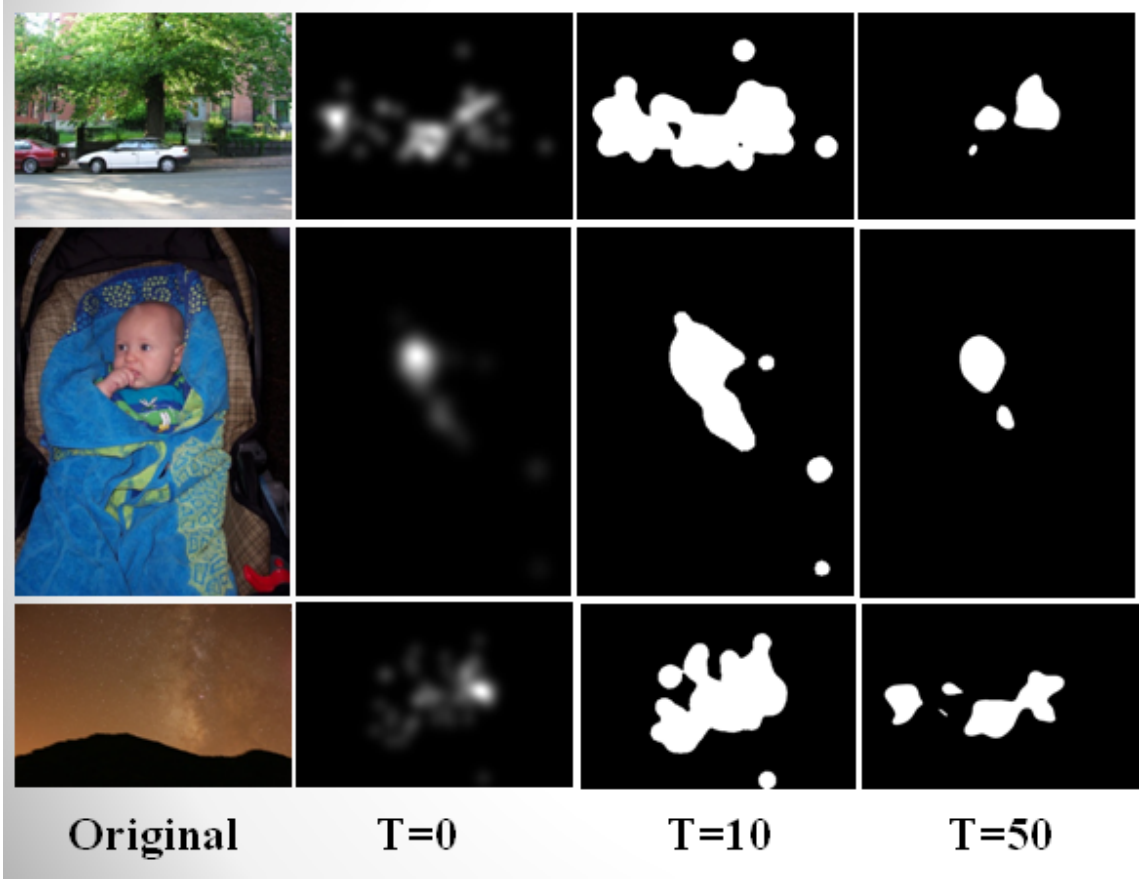

Figure 2.2: Examples of quasi-objects carried out using the Ramik's algorithm showing input images (column a), the corresponding eye-fixation map (column b) and matched quasi-objects obtained setting $T = 10$ (column c) and $T = 50$ (column d)

in a more general way, an appropriate tuning of the saliency detection parameters allows bringing closer the saliency detection process and the human-like observation of the same vista.

2.3.1.2 Object Detection Algorithm as Quasi-eye Fixation Algorithm

According to [Kienzle 09], a center-surround feature based algorithm can be used in eye fixation prediction tasks. Let us interpret saliency map in eye fixation problem (eye fixation saliency map, EFSM) as a map, which shows a distribution of image's pixels' likelihood to match up pieces of images' items which may draw an attention. In this case we can say, that saliency maps, obtained via object detection algorithm, can be interpreted as EFSM, and can be compared to eye fixation maps from MIT1003 database. This work mode doesn't need steps 3–4 (segmentation map and object detection).

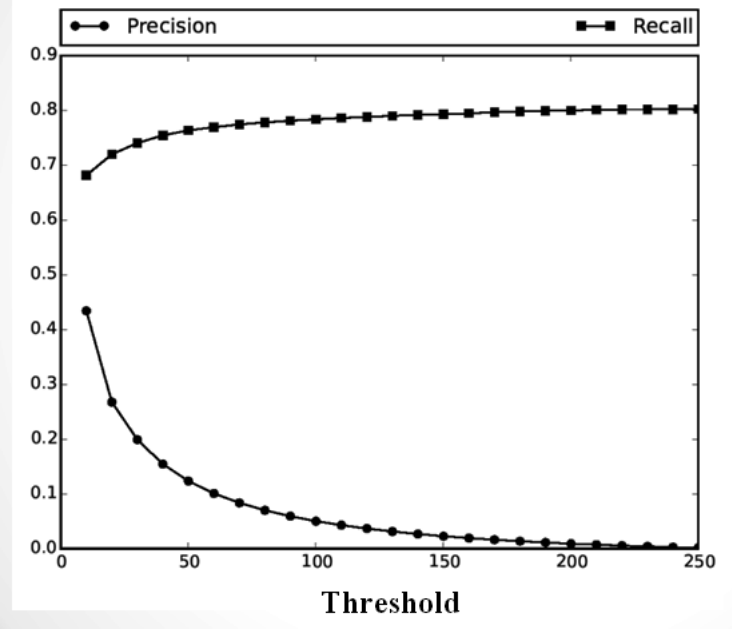

Figure 2.3: Precision and Recall average values over all 1003 images versus different salient region threshold values $T \in [10, 250]$.

Equating the final saliency map to an EFSM, the following question can be formulated: "how similar are the final saliency map and the human's eye fixation behaviour?". In other words, could the salient objects' detection algorithm fit human-like visual attention?

The answer to this question could be loomed by evaluating, through the MIT1003 database's images and the corresponding actual eye-fixation maps held by this database, the likeness between eye fixation map and final saliency map. As shown previously in section 1.3, we should use at least 5 indicators to be able to assess a model fairly.

In Table 2.1 shows some evaluation scores of Ramik's algorithm. We have run it with different values of $\{IRC; WSC\}$ parameters (where IRC, Image Resize Coefficient represents an inner technical parameter of pre-scaling for the input image, applied automatically in order to decrease the computational complexity while not losing too much in performance – more on the subject in subsection 4.2.1), and the best scores were achieved with different set of parameters (for example, $\{IRC = 0.25; WSC = 0.15\}$ or $\{IRC = 0.2; WSC = 0.14\}$). Yet, the default set of values $\{IRC = 0.2; WSC = 0.4\}$, as given in [Ramík 12], yields scores poorer, than some other sets.

The indicators' scores, obtained within quite a particular setting-option holding the WSC parameter constant for all images in dataset, highlight three outcrops. The first one is that some choices of the parameters (namely $WSC = 0.14$ and

Table 2.1: Examples of evaluation scores' values for four different selection options of the parameters of Ramik's algorithm

$IRC; WSC$	AUC_{Judd}	AUC_{Borji}	$sAUC$	NSS	KL
0.15; 0.1	0.6902	0.6506	0.6607	0.7203	0.7537
0.2; 0.14	0.7286	0.6733	0.6921	0.7531	0.7374
0.2; 0.4	0.7246	0.6524	0.6513	0.7352	0.7603
0.25; 0.15	0.7286	0.6749	0.6831	0.7597	0.7379

$WSC = 0.15$ while $IRC = 0.2$ and $IRC = 0.25$) conduct to better matching of the human-like way of gazing the images of the considered MIT1003 database. The second one relates AUC-like indicators' behaviour versus the parameters tweaking, showing that those same choices lead to an effectual enhancement of the matching algorithm's propinquity with human's eye fixation mechanism: in other words, it is not an accidentally emergence of some human-like visual behaviour. The last remark joins the concluding statement of subsubsection 2.3.1.1: an appropriate tuning of the saliency detection parameters allows bringing closer the saliency detection process and the human way of observation.

However, an additional overall remark relating the obtained scores steers to be conscious that an accurate tuning of the model will require the fine-tuning of all additional parameters embroiled in other stages of the investigated system.

2.4 Top-Down Unit

As TD models generally are classified into three different types (as mentioned in subsection 1.2.3) which are pretty difficult to combine, we need to narrow down the scope by several general restrictions and conditions:

- In general, such a model have to be able to work in real-time (up to 1 sec. in optimal implementation, with possible iterations of state "robot is thinking", up to 5-10 seconds);
- This model can consist of several other models combined;
- Model combination method should be able to change its parameters depending on task;
- The default task is "free visual search", as a natural human behaviour;
- Other possible general tasks might be defined as "free visual search with additional attention to known objects", "search for target object", "search for target pattern", "reconnaissance of unknown scene".

Thus we can define several parts of this approach which we can call as a TD visual attention model, because it creates visual attention prediction based on cognitive phenomena, and in general combining two of three main types of TD models, namely it is a visual search model with a hint of task drive.

As it has been already stated, in top-down direction the attention is spanned towards objects with additional characteristics, viewed and provided by higher cognitive abilities; in such a manner, one of the fundamental higher cognitive ability is an ability to detect and recognize the object. It is pertinent to note, that this task is not equal to the standard "salient object detection problem" (as outlined in subsection 1.2.1) per se, but is a complex task of both detection and recognition.

The recognition module in this case is the one that provides both the answers to the questions "*Is there any recognizable object in the scene?*" and "*If yes, then what is this object, and what is your confidence level for this act of recognition?*".

But there is also a necessity to answer another question, – "*If there is an object previously learnt, how do we recognize it?*".

In order to answer this question, we have to not only use object detection and recognition as main tool in all TD attention, but also use bi-directional information exchange between TD model and memory unit in the principal scheme, – whether the model memorizes an object, the recognition module should be able to recognize it.

It would be too time-consuming to create new approaches for each step, rather than choose an existing one, while there exist many works, spanning through decades, both on recognition and memory models; also, such work would be out of a timespan of a three year doctoral thesis.

From another point of view, it is possible to take either the several state-of-art algorithms and to compare them, or to find several existing comparison works and to verify the existing judgements, in order to make a choice of the most usable algorithms in recognition and memory modeling. As shown in subsubsection 1.4.1.3 and already discussed in section 2.1, we can let us use pretty time-consuming models in this part of the bigger system, without fear of losing the quasi-real-time efficiency. Thus, the robot is still able to be autonomous: without any network connection (if it is lost), it still can use the basic, bottom-up part of the model.

2.5 Memory and Decision Unit

Based on some rule and/or decision making module, a "mix" of attention maps, produced by TD and BU units, should be done. As a result, we would have a more

weighted saliency map, produced by both types of attention direction, along with complimentary results, such as “Representative points” (produced by bottom up model extension), or the recognition data (produced by top down model).

While the name of this module stands for two pretty broad and general concepts, it is used in more narrow sense – we need to store seen objects (”short term memory”), to be able to recognize them in future and be able to apply a correct label onto these objects, and make the decisions based on the information gathered previously – so this is more just a question of data systematization.

In contemporary cognitive psychology the defining model of short-term memory changed through time: the keystone Atkinson-Schiffrin model [Atkinson 68], which implies division of the model into modal modules (also known as multi-store model or modal model), evolved in time into Baddeley-Hitch model of ”working memory” [Baddeley 74] (additionally modified by Baddeley in [Baddeley 00]).

There have been several approaches to implement ”working memory” in robotics, namely [Skubic 04] or [Phillips 05]. While Skubic’s multi-agent approach to STM and LTM has shown very good results, in our situation it represents too high complexity; for our purposes of storage of objects with images, or concepts of such objects, a simplified Baddeley model can be repurposed (model shown in Figure 2.4).

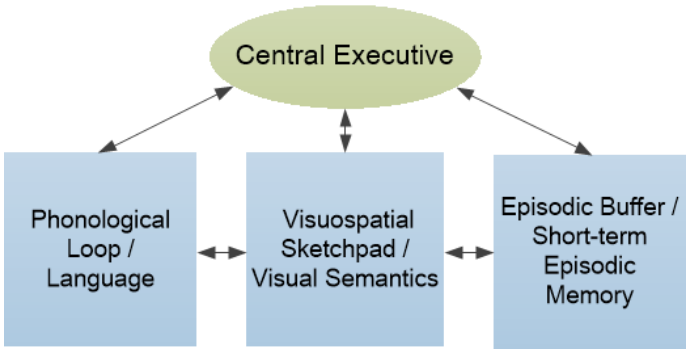

Figure 2.4: Baddeley-Hitch working memory model as contemporary cognitive psychology view on STM and LTM mediator

While such model applies to the whole sensory input, including audio, tactile, navigation ”senses” of human (or robot), in order to apply it only to visual memory we can simplify or repurpose its parts, using Baddeley’s STM as a general basic structure.

Thus, ”Phonological Loop” originally represents audio input coherent with visual input; in our case, we can repurpose it as a part where input task, or input data from recognition module is analysed in terms of language; e.g., a part which should answer a question *” If we need to find object **A**, but we found object **B**, whether it is significant?”*, or if re-phrased, what is a quantitative estimation of the closeness

between objects **A**and **B** and whether the robot should react?

The "Visuospatial Sketchpad" implies to existence of navigation sensor ("inner compass" along with vestibular system of human); while the question of navigation has always been an outstanding problem in modern robotics, this question is out of the scope of this work. Thus, we can repurpose it as a map of visual knowledge, which acquires data from recognition module and stores it.

The "Episodic Buffer" is assumed to be a limited-capacity temporary storage system that is capable of integrating information from a variety of sources. It is assumed to be controlled by the central executive, which is capable of retrieving information from the store in the form of conscious awareness, of reflecting on that information and, where necessary, manipulating and modifying it. The buffer is episodic in the sense that it holds episodes whereby information is integrated across space and potentially extended across time. As in context of our research, it can be represented as a constantly changing self-organizing model, namely Kohonen map, which represents another mean of estimating similarity between freshly acquired sensory information and several previously gathered images.

And as for "Central Executive", – it represents the part which distributes the energy between the other memory parts, along with the information redistribution. In our case this should be a submodule which queries all the parts of memory, and inducts the decisions over all the gathered data.

2.6 Conclusion

This chapter discusses general outline of a combined visual attention model, which not only combines two different directions of visual attention via several bio-inspired techniques, but also provides additional recognition along with some importance estimation of the visual data, allowing the mobile robot to react in a more informative fashion. Such a system is described in general manner; more details on it are given in chapter 3.

By applying the described system to several specific datasets in order to obtain validation, we can show the usability of given sets of parameters along with the whole system efficiency as well; more on the topic of validation of the system in simulations and the real-world tasks will be discussed in chapter 4.

3 | Theoretical Basis of the Combined Visual Attention Model

3.1 Introduction

As we have discussed and stated previously in chapter 2, visual perception models, usable in autonomous robotics, should have higher operational speed, and this condition provides several implementation constraints onto them nowadays: the embedded part of such a model would have a "bottom-up" direction of a visual scene study, because "top-down" model can not be considered as "fast" due to large resource demand.

Focusing on the dilemma of visual attention for autonomous visual perception, in this chapter we propose a more in-depth view on the general model for combined artificial visual attention, along with its parts:

- Visual-Attention-based Autonomous Artificial Vision (VA^3V) as bottom-up unit, along with its genetic tuning process. Statistical foundation and bottom-up nature of the proposed model provide the advantage to make it usable without prior information, making it suitable for usage in autonomous artificial vision in general.

- Set of object recognition approaches as top-down unit. Mostly the choice of techniques was done in Chapters 1 and 2, so in this chapter we will consider some practical points of implementation.

- Memory and decision unit as set of modules incorporated into schema of Baddeley-Hitch short-time memory.

In section 3.2 we introduce the groundwork of the proposed BU approach. In the section a statistical foundation of visual saliency is presented. Then section 3.3 provides description of the proposed approach as an extension of existing techniques, fine-tuned by an evolutionary process along with experimental results on MIT1003 and Toronto image datasets, comparison to currently best algorithms used in the

aforementioned field. The reported results show evaluation scores comparable to the state-of-art algorithms with advantage of a comprehensive solid theoretical basement and fast execution speed of our approach. The section is concluded by demonstration of a further extension to this model in terms of expanding visual attention similarity and saccade prediction. Then, section 3.4 provides some in-depth vision over other two parts of the general system, TD unit and decision unit, and then the chapter is concluded with sketching further perspectives of the presented model.

3.2 Theoretical Basis of Statistically Driven Artificial Visual Attention and the Proposed Approach

The proposed approach uses some concepts of previously mentioned Ramik's algorithm (see subsection 2.3.1 and [Ramík 12]), as these concepts have been justified to be usable in eye fixation problem.

Based on saliency detection concept, the proposed Visual-Attention-based Autonomous Artificial Vision (VA^3V) approach gets underway the integration of human-like bottom-up visual attention by launching an eyes fixation mechanism based tuning of the saliency detection process. This is done through a Genetic Algorithm (GA) based evolutionary process shoving the saliency detection toward the human-like eyes-fixation behaviour.

Figure 3.1 shows the operational block diagram of the proposed approach. As it could be noticed from this figure, the two first blocks are liable for working out two kinds of saliency maps, representing two different saliency levels: "Global Saliency Map" (GSM) and "Local Saliency Map" (LSM). These two parts can work in parallel, as they represent two different algorithms, applied to the same input. The results are two maps, which are then taken as input for computation of the final saliency map. The next block, labeled "Fusion – Final Map", is devoted to these fusion operations.

The first fusion operation combines the two above-mentioned saliency maps in order to carry out the final saliency map. The second one performs a weighted combination of GSM and a set of LSMs.

Finally the last block "Gaussian Blob Filter" applies several filters (to be explained in details in following subsections) onto the final saliency map leading to the "Visual Attention Map" which holds compulsory features relating relevant visual information of the perceived scene (image). Then, this key-map might be used either

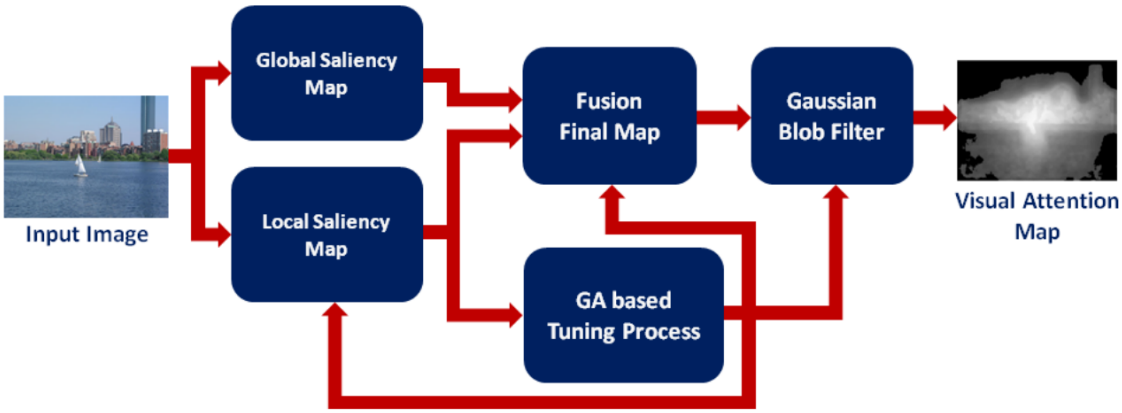

Figure 3.1: Operational block diagram of the proposed VA^3V model

for extracting the outstanding items (e.g. objects' detection recognition tasks) or for acting as a visual control loop controlling a machine's (namely robot's) actions.

The aforementioned GA-based tuning process is depicted on the diagram as another block, which influences operations in three parts (in saliency computation, in fusion and in filtering). This shows the idea that the tuning process creates various involved parameters, which are used throughout the whole operational flow.

It is also pertinent to emphasize that the investigated system admits two operational modes: the "tuning mode", which acts as some kind of learning process, with input in form of the training datasets with corresponding "ground truths", and output in form of a set of suitable parameters to be used in the system; and the "operating mode" carrying out the so-called visual attention map once the system's parameters have appropriately been tuned.

3.2.1 Saliency Map Computation

As a start, we need to formally define all the elements of the problem's input.

Let us suppose the image $\Omega_{\mathrm{RGB}}(x)$, represented by its pixels $\Omega_{\mathrm{RGB}}(x)$ in **RGB** color space, where $x \in \Psi^2$ denotes 2d pixel position. Let $\Omega_{\mathrm{R}}(x)$, $\Omega_{\mathrm{G}}(x)$ and $\Omega_{\mathrm{B}}(x)$ be the colors values in channels R, G and B, respectively. Similarly, we can define the image in other color spaces (e.g., **YCrCb**) – $\Omega_{\mathrm{YCC}}(x)$, $\Omega_{\mathrm{Y}}(x)$, $\Omega_{\mathrm{Cr}}(x)$ and $\Omega_{\mathrm{Cb}}(x)$ in **YCrCb** color space. Finally, let $\overline{\Omega}_{\mathrm{R}}$, $\overline{\Omega}_{\mathrm{G}}$ and $\overline{\Omega}_{\mathrm{B}}$ (or $\overline{\Omega}_{\mathrm{Y}}$, $\overline{\Omega}_{\mathrm{Cr}}$ and $\overline{\Omega}_{\mathrm{Cb}}$) be median values for each channel throughout the whole image.

As mentioned at the beginning of the present section, two kinds of saliency maps, representing two different saliency levels, are created: "Global Saliency Map" (GSM) and "Local Saliency Map" (LSM). The GSM handles the ability to catch attention dealing with the general contrast of all and any part of the image, while LSM conveys

fine levels analysis dealing with the local center-surround contrast in the considered image. The computation of the saliency map is in fact the computation of these so-called GSM and LSM.

Global Saliency Map – also called GSM, denoted as $M_{\mathrm{G}}(x)$, – is a result of non-linear fusion of two elementary maps $M_{\mathrm{Y}}(x)$ and $M_{\mathrm{CrCb}}(x)$, relating luminance and chromaticity separately. Equations (3.1) – (3.3) detail the calculation of each elementary map as well as the resulting GSM.

The Equation 3.1 depicts calculation of elementary map over Y-channel, where each value for position x is calculated as modulus of difference of the corresponding value in $\Omega_{\mathrm{Y}}(x)$ and median value for this channel. The Equation 3.2 does similar calculation by combining such differences of two channels, **Cr** and **Cb** in "Euclidean distance" fashion.

$$M_{\mathrm{Y}}(x) = |\overline{\Omega}_{\mathrm{Y}} - \Omega_{\mathrm{Y}}(x)| \tag{3.1}$$

$$M_{\mathrm{CrCb}}(x) = \sqrt{\left(\overline{\Omega}_{\mathrm{Cr}} - \Omega_{\mathrm{Cr}}(x)\right)^2 + \left(\overline{\Omega}_{\mathrm{Cb}} - \Omega_{\mathrm{Cb}}(x)\right)^2} \tag{3.2}$$

$$M_{\mathrm{G}}(x) = \frac{1}{1+\mathrm{e}^{-C(x)}} M_{\mathrm{CrCb}}(x) + \left(1 - \frac{1}{1+\mathrm{e}^{-C(x)}}\right) M_{\mathrm{Y}}(x) \tag{3.3}$$

As blending function for the composite "global saliency map" $M_{\mathrm{G}}(x)$ we use the logistic sigmoid. This blending of the two elementary saliency maps together in Equation 3.3 is driven by a function of color saturation of each pixel. For this purpose, the color saturation C_{c} is defined, and is calculated from **RGB** color model for each pixel as pseudo-norm, normalized to 0–1 range (as shown in Equation 3.4 by division to 255, where 255 is maximal value in any of R, G, B channels). When C_{c} is low (too dull, unsaturated colors), more importance is given to intensity saliency $M_{\mathrm{Y}}(x)$. When C_{c} is high (vivid colors), chromatic saliency $M_{\mathrm{CrCb}}(x)$ is emphasized.

The values in Equation 3.5 are chosen in order to fit the logistic sigmoid [Ramík 12, p. 63].

$$C_{\mathrm{c}}(x) = \frac{\left(\max\left(\Omega_{\mathrm{R}}, \Omega_{\mathrm{G}}(x), \Omega_{\mathrm{B}}(x)\right) - \min\left(\Omega_{\mathrm{R}}(x), \Omega_{\mathrm{G}}(x), \Omega_{\mathrm{B}}(x)\right)\right)}{255} \tag{3.4}$$

$$C(x) = 10\left(C_{\mathrm{c}}(x) - 0.5\right) \tag{3.5}$$

Local Saliency Map – also called LSM, – is a map, calculated on the basis of idea of center-surround histograms (initially proposed in [Liu 11]), involving statistical properties of two centered windows (over each pixel) sliding alongside whole the

image. Based on this concept the LSM is obtained from a non-linear fusion of two statistical properties-based maps (same mechanism as in the GSM calculation), relating luminance and chromaticity of the image.

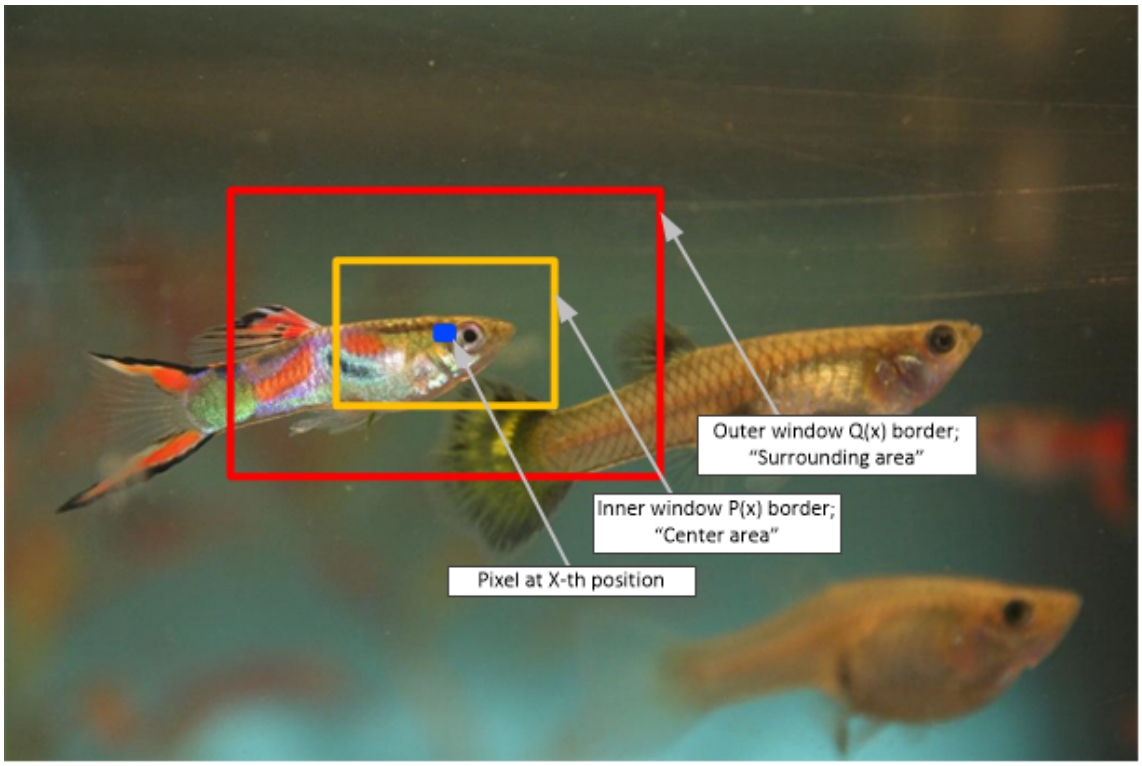

Figure 3.2: Illustrative example of areas $P(x)$ and $Q(x)$, representing center-surround antagonism

Let us suppose a sliding window $P(x)$ of size p, centered over a pixel x, which is compared to its surrounding area $Q(x)$, so that $(Q(x) - P(x)) = p^2$ (as illustrated in Figure 3.2). Let a center histogram $H_{\mathrm{C}}^{ch}(x)$ be a histogram of pixel-applied intensities in each ch – channel in **YCrCb** space, – thus speaking about three histograms, one for each channel, in window $P(x)$ (with $H_{\mathrm{C}}^{ch}(x,i)$ being a value of i-th bin of this histogram, respectively), and a surrounding histogram H_{S}^{ch} — a histogram of pixel intensities in area $Q(x)$. Then the center-surround feature $d_{ch}(x)$ can be calculated for each channel $ch \in \{\mathrm{Y}, \mathrm{Cr}, \mathrm{Cb}\}$ as in (3.6) — a normalized sum of difference over all 256 histogram bins. For pixels near the edge of image the areas $P(x)$ and $Q(x)$ are only over the image itself.

$$d_{ch}(x) = \sum_{i=1}^{255} \left(\frac{H_{\mathrm{C}}^{ch}(x,i)}{|H_{\mathrm{C}}^{ch}(x)|} - \frac{H_{\mathrm{S}}^{ch}(x,i)}{|H_{\mathrm{S}}^{ch}(x)|} \right) \tag{3.6}$$

In Equation 3.6 $|H_{\mathrm{C}}^{ch}|$ and $|H_{\mathrm{S}}^{ch}|$ represent sum over all histogram bins.

The LSM (denoted as $M_{\mathrm{L}}(x)$), resulting from a non-linear fusion of the so-called

center-surround features (denoted as $d_{ch}(x)$), is calculated in Equation 3.7, with coefficient $C_\mu(x)$ being an average color saturation over window $P(x)$. It is calculated in Equation 3.8 as an average of saturation $C(x)$ on area $P(x)$ from Equation 3.5.

$$M_{\mathrm{L}}(x) = \frac{1}{1-\mathrm{e}^{-C_\mu(x)}} d_{\mathrm{Y}}(x) + \left(1 - \frac{1}{1-\mathrm{e}^{-C_\mu(x)}}\right) \max\left(d_{\mathrm{Cb}}(x), d_{\mathrm{Cr}}(x)\right) \tag{3.7}$$

$$C_\mu(x) = \frac{\sum_k^{P(x)} C(k)}{p} \tag{3.8}$$

Let us return back to the parameter p, introduced at the beginning of the present subsection, which regards the choice of the center-surrounding windows' sizes (e.g. satisfying the condition $Q-P=p^2$ in general; and for the "border" pixels, where the sliding window would be out of image margins, the condition is interpreted as $Q-P \leq p^2$). We need to note that small values of p direct the saliency detection process toward highlighting details or small items, while larger values of this parameter make the saliency detection process stressing bigger items of the image.

To formulate the task of parameter p, let $n \times m$ be the image's size, where n and m are numbers of horizontal and vertical pixels of image, respectively. Q may be seen as a fraction of the image and be expressed as $Q = \alpha\, n \times m$, where $0 \leq \alpha \leq 1$. α may be interpreted as the parameter linking to the visual attention's perimeter. Within such statements, we can define the ratio $\frac{P}{Q}$ that we denote as γ ($0 \leq \gamma \leq 1$), which could be seen as the visual attention grain, representing the attention scale. The area p^2, corresponding to the surrounding part of window $Q(x)$ – the $Q-P$ part, also a fraction of the image, – could be expressed as $p^2 = \alpha(1-\gamma)n \times m$, showing the parameter p as a function of γ, α, and the image sizes:

$$p = \sqrt{\alpha(1-\gamma)n \times m} \tag{3.9}$$

Therefore, p could be interpreted as the parameter relating the "visual attentiveness scale" of the system (the parameter controlling the system's visual attentiveness), depending both on the considered "visual attention's perimeter" Q (involved through α), and on "visual-attention grain" $\frac{P}{Q}$ (involved through γ).

An appealing way of handling the choice of the parameter p (and also the sizes of windows P and Q) is to be able to define it by only one coefficient – window size coefficient, WSC, which by itself should not be dependent on the size of image, but should provide both the scale and the grain of visual attention:

$$p^2 = WSC^2 \times n \times m \tag{3.10}$$

$$WSC = \sqrt{\alpha(1-\gamma)} \tag{3.11}$$

It has been shown that the "equilibrium" value of relative size between "center" and "surround" windows can be $\gamma = \frac{1}{2}$ [Liu 11]. In this case window size coefficient defines both the sliding windows definitively:

$$P(x) = WSC^2 \times n \times m \tag{3.12}$$

$$Q(x) = 2WSC^2 \times n \times m \tag{3.13}$$

Final Saliency Map – or just Saliency Map (SM), denoted as $M_{\text{final}}(x)$, is a map resulting from a non-linear fusion of GSM and LSM as shown in (3.14).

$$M_{\text{final}} = \begin{cases} M_{\text{L}}(x) & \text{if } M_{\text{L}}(x) > M_{\text{G}}(x) \\ \sqrt{M_{\text{L}}(x) M_{\text{G}}(x)} & \text{otherwise} \end{cases} \tag{3.14}$$

3.3 VA^3V model

3.3.1 Fusion and Gaussian-Blob-Based Adaptive Filtering

Referring to Figure 3.1, two kinds of fusions are carried out by the "Fusion — Final Map" block. The first one has already been shown at the very end of subsection 3.2.1, which is a nonlinear fusion of GSM and LSM leading to the SM (final saliency map of Ramik's algorithm). This fusion concerns as well the tuning mode as the operation mode. The second one, performing a weighted fusion, combines GSM and a set of LSMs. This fusion operation is concerned to be able to apply several visual attention grains onto the input image, because a fusion of different LSMs gives better overall results, even if they are fused in a simple linear matter, as shown in Equation 3.15:

$$M_{\text{E}}(x) = w_0 * M_{\text{G}}(x) + \sum_{i=1}^{N} w_i * M_{\text{L}\,i}(x), \qquad \sum_{i=0}^{N} w_i = 1. \tag{3.15}$$

Where N represents number of LSMs to be fused, and w_i, $(i \in 0..N)$ are complementary weight coefficients of the fusion process. Values of N and w_i are parameters and subject to a tuning process, as it is a question of resources and performance – the more fused LSMs we have, the better is the result, but the higher is the time

spent to the processing.

The resulting map, also a saliency map, is then filtered by a "Gaussian-Blob Filter" and – in tuning mode – used by GA-based tuning process for tuning various system's parameters in order to shove the saliency detection process toward the eye-fixation mechanisms.

The Gaussian-Blob-based filtering, suggested and discussed in several works relating the eye fixation problem (e.g., [Bruce 07], [Riche 13a], [Judd 09] and [Tatler 07]), and based on humanitarian concept of prevalence of centered attention over peripheral interest, consists in sifting the above-mentioned final saliency map by a 2-dimensional Gaussian center blob which profile is defined by Equations (3.16)-(3.18).

$$G(i,j) = A \times \exp\left(-\frac{(i-i_0)^2}{2\sigma_i^2} - \frac{(j-j_0)^2}{2\sigma_j^2}\right) \tag{3.16}$$

$$\sigma_i = \frac{\mathrm{n}}{2 * Nar} \qquad and \qquad \sigma_j = \frac{\mathrm{m}}{2 * Nar} \tag{3.17}$$

$$\widehat{M_{\mathrm{E}}}(i,j) = w_{\mathrm{G}} * G(i,j) + w_{\mathrm{E}} * M_{\mathrm{E}}(i,j) \tag{3.18}$$

Here pixel x is represented by its respective coordinates (i,j), and (i_0,j_0) are coordinates of the image's geometrical center (e.g. center of the concerned picture). $sigma_i$ and $sigma_j$ relate the Gaussian blob's radius, A represents amplitude, Nar is a scale parameter which represents "narrowness" of the blob, and (as previously already defined) n and m are number of pixels in the image horizontally and vertically, respectively.

An appealing way of tuning $sigma_i$ and $sigma_j$ is to set them equal $\sigma = \sigma_i = \sigma_j$ as $\sigma = \frac{\sqrt{\mathrm{n}\times\mathrm{m}}}{2*Nar}$. This leads to shaping a symmetrical Gaussian blob, decreasing computational time.

$$M_{\mathrm{VAM}}(x) = \begin{cases} \widehat{M_{\mathrm{E}}}(x) & \text{if } \widehat{M_{\mathrm{E}}}(x) > FT \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

The filtering operation itself consists of a weighted boosting of the map resulting from the fusion block as depicted in Equation 3.18. The final map, namely the Visual Attention Map (VAM), results from an adaptive threshold-based purging following Equation 3.19, operated on the map issued from Gaussian operation. FT here represents some "final threshold", – parameter, whose value is subject to tuning process (along with the Gaussian weighted sum's coefficients w_{G} and w_{E}, and the parameters of Gaussian blob itself – A and K).

Time efficiency of such model heavily depends on the number and sizes of sliding windows in local saliency map calculation. Therefore algorithm time complexity (and time consumption) depends on the image size: so let us denote the size of image in the moment of processing as $N = n \times m$ "square pixels".

If we use a naïve approach ("sliding window" loop over each pixel in bigger loop, calculating each histogram "from scratch" in each iteration), local saliency map algorithm complexity can be estimated as $O(N^2)$; But if we remember the histograms of previous iteration and just apply the difference on each iteration, the algorithm complexity estimation in this part can be dropped to $O(N * \log N)$, while each other step (global saliency map, ensemble fusion, Gaussian blob, final threshold) represents simple $O(N)$ complexity, thus giving us $O(N * \log(N) + N)$ worst complexity.

The complexity analysis for CNN-based approaches in visual saliency had never been done; yet, there exist several works about complexity of general CNNs, used mostly in object recognition (such as ResNet evaluation [He 15] or general CNN evaluation [Gouk 14]). According to given works, the worst time complexity evaluation (after tweaking) for a CNN is $O(M * P * I + P * I \log I + M * I * \log I)$, where P is the number of input feature maps, M is the number of output feature maps, I is the number of elements in each input map. As, by design, usually the number of feature maps is constant for each CNN architecture (yet significantly bigger than 1), and I represents the same number of square pixels as N, we can state that $O(N * \log(N) + N) \leq O(M * P * I + P * I \log I + M * I * \log I)$, giving us the theoretic justification of better time efficiency of VA^3V approach, if comparing with CNN-based approaches (eDN, SalGAN, DeepGaze II, etc.)

3.3.2 GA-Based Evolutionary Tuning Process

The investigated VA^3V system takes advantage from the illustrated analogy by fitting in human-like conduct into the system through using the saliency detection process as the eye fixation mechanism. This adaptation is achieved through GA-based evolutionary fine-tuning of various parameters, controlling different computational blocs, impacting the three main computational levels of the system:

- namely, the saliency maps computation via WSC, N and w_i parameters adjustment;
- the filtering process by the Gaussian blob-shape's parameters' regulation;
- and the final matching through the FT threshold adjustment.

The choice of GA-based evolutionary process is motivated by contribution of a quite large amount of interdependent parameters and the requirement of a reciprocal tuning of these parameters regarding the involved processes. The real human gaze depends on a big set of hardly formalized factors [Itti 07], such as: is it a free looking task or a visual search for an exact object; what is the amount of given time; if exists the a priori knowledge, etc.

According to the practical task needed to be solved by this algorithm, one of the several sets of parameters should be chosen based on input data characteristics. Due to this, the fact of considering the whole set of the involved parameters as a genome, offers an appealing way of their optimization within the frame that privileges the global effect of whole the set versus the individual impact of each of those parameters. Taking into account the expected target (e.g. incorporation of human's vision skills), the fitness function is based on the previously described quality indicators (see section 1.3) – 3 different AUC-based metrics, K ullback–Leibler divergence and Normalized Scanpath Saliency, – as it has been established that these indicators are not fully correlated between themselves, thus giving us an appealing possibility to use a combined, "integral" metric of model quality, as some kind of a sum (or weighted sum) of all afore-mentioned metrics.

To do this, we need to define explicitly several concepts of such a metric:

- All components, used in such a sum, should be scaled in a comparable way. If AUC-based metrics are scaled from 0 to 1, and the higher is the value – the better is the model, the KL-divergence is controversial to them (the lower is the indicator, the better is the model);
- The NSS indicator shows similar to AUC-based metrics dynamic. As it is outlined in [Bylinskii 16], NSS might be the most important metric, it can be wise not to decrease its importance, and allow it to have the scaling up to ideal score of ≈ 3.2;
- The really bad, unfitting values of indicators (namely, if the model performs worse, than chance) should be additionally penalized in order not to propagate bad parameter sets in generations due to compensation by other good results.

According to these concepts, we can formulate an "integral" metric of quality which could also be defined as a fitness function of a genetic model, as given by equations (3.20)–(3.22), where N represents a number of images in dataset, used for tuning, and i is iterator over all these images.

$$Fit = \sum_{i=1}^{N} (AJ(i) + AUC_{\text{Borji}}(i) + sAUC(i) + NSS(i) + KL(i)) \quad (3.20)$$

$$AJ(i) = \begin{cases} AUC_{\text{Judd}}(i) & \text{if } AUC_{\text{Judd}}(i) > 0.5 \\ AUC_{\text{Judd}}(i) - 1 & \text{otherwise} \end{cases} \quad (3.21)$$

$$KL(i) = 1 - KL_{\text{div}}(i), \quad (3.22)$$

As shown in Equation 3.21, $AJ(i)$ is one of the penalty achieving components – if $AUC_{\text{Judd}}(i)$ represents poor performance, namely being lower than 0.5, which means performing worse than chance, – the overall fitness function is penalized by 1 (for each poorly processed image). Another penalty component is imposed on divergence $KL(i)$: if real divergence is higher, than 1, – the fitness function is also penalized due to a negative value.

In case of $AUC_{\text{Borji}}(i)$ and $sAUC(i)$ metrics, they do not have exact "chance" threshold due to their "randomized" nature by definition. And NSS metric has a possibility to be negative by design in case of poor performance, so that we do not need to apply additional constraints or penalization components here.

The whole tuning algorithm in this case represents an usual genetic approach. A chromosome represents set of parameters, which define one VA^3V model, and is shown in Table 3.1.

As all items in the chromosome are either integer or decimal, it is not possible to apply standard GA-procedures of mutation and crossover in their initial, binary form [Holland 92]. Due to this we must use weaker definition of these two procedures, where ***chromosome mutation*** is defined as *random decimal reinitialization of one of the chromosome parameters*, and ***chromosomes' crossover*** is *an exchange of several parameters, chosen randomly, in whole.*

3.3.3 Tuning Viability Regarding Likeness with Human-like Vision

First question about the tuning process regards the general possibility of such a process to tune a randomly initialized VA^3V model to represent a human-like vision. For this, by referring to the dual analogy discussed in subsection 2.3.1, the so-called visual attention map (resulting from the investigated system) is compared against predicted eye fixation maps obtained from the best eye fixation prediction algorithms, as referenced in subsection 1.2.2: namely SalGAN, eDN, BMS and RARE2012 (DSCLRCN algorithm authors did not disclose any implementation to

Table 3.1: Chromosome, consisiting of VA^3V model parameters tuned by genetic approach

Parameter	Definition	Possible value
IRC	image resize coefficient for the initial scaling of image	*decimal, from 0 to 1*
N	number of local saliency maps calculated	*integer, from 1 to 4*
WSC_i, $(i \in [1..N])$	window size coefficients, which define the p-parameter of each local saliency map and, therefore, the scale and grain of local saliency	*N decimals, from 0 to 1 each*
w_i, $(i \in [1..N])$	weighted sum coefficients which define the local & global saliency maps' fusion	*N decimals, from 0 to 1 each*
A	amplitude of 2-dimensional Gaussian blob intensity	*integer, from 1 to 255*
Nar	scale factor, "narrowness" of 2-dimensional Gaussian blob	*decimal, from 0 to +*inf
w_G	Gauss map wight factor in weighted fusion	*decimal, from 0 to 1*
FT	final threshold	*integer, from 0 to 255*

be used in arbitrary assessment).

In fact, the assessment protocol is based on the following reasoning: ifm as stated before, the above-mentioned state-of-the-art algorithms are the best ones in modeling the aforementioned human's visual skill (e.g. the eye fixation mechanism), then the evaluation of the investigated system versus those best algorithms will reflect its quality regarding the eye fixation mechanism and thus will reveal the likeness of the VA^3V system and the human-like vision.

Pursuing the same reasoning, this evaluation will also reflect the tuning process' viability. Such test has been done by using both MIT1003 and TORONTO benchmark databases relating the eye fixation modeling. while we have to say that there are several other available benchmark databases (as: FiWI dataset dedicated to webpage saliency presented in [Shen 14] or EyeCrowd dataset devoted to face recognition presented in [Jiang 14]), our choice of using the two above-mentioned has been motivated by the fact that they include "ground truth" eye fixation maps obtained experimentally by involving humans and thus truly representative of eye fixation mechanism. The evaluation has been performed through all 5 indicators previously presented. The obtained results are summarized and presented as scores in Table 3.2.

Examples of visual attention maps carried out by VA^3V system and predicted

Table 3.2: Comparison with state-of-art algorithms on MIT1003 and Toronto datasets

Algorithm	Dataset	AUC_J	AUC_B	KL	NSS	$sAUC$	Fit
eDN		0.8501	**0.7685**	0.6681	1.2969	0.5834	N/A
BMS		0.7806	0.6103	0.6261	1.2524	0.5837	N/A
RARE2012	MIT1003	0.7845	0.6171	0.6052	1.3215	0.5792	N/A
SalGAN		**0.8711**	0.6949	**0.3741**	**2.1931**	0.6605	N/A
VA^3V		0.8322	0.7445	0.6389	1.2021	**0.7658**	3.4696
eDN		0.8541	0.6291	0.4808	1.581	**0.6999**	N/A
BMS		0.7461	0.5368	0.4201	2.1914	0.6366	N/A
RARE2012	TORONTO	0.7688	0.5381	0.3910	2.2785	0.6255	N/A
SalGAN		**0.8632**	0.6123	**0.3692**	**2.341**	0.6954	N/A
VA^3V		0.8372	**0.6375**	0.4504	1.6567	0.6448	4.3255

eye fixation maps achieved by the above-mentioned leading algorithms are provided by Figure 3.3. The first and last columns in Figure 3.3 give input stimulus (e.g. images) and the corresponding grand true (e.g. experimental) eye-fixation maps, respectively.

The Table 3.2 highlights several appealing features. The first one relates the fact that VA^3V scores are comparable with those of the above-mentioned state-of-art best algorithms. The second remark goes to the fact that VA^3V scores are even higher than those of two among the four algorithms, ranking it clearly as comparable to the "leading" algorithms (e.g. eDN and SalGAN). However, we need to note that the prevailing position within this group of leading algorithms is a result of the GA-based tuning of VA^3V system's parameters, reinforcing the importance of such built-in evolutionary tuning mechanism.

Moreover, the closeness of scores to those of the best algorithms' performances, visibly, points up the emergence of an eye-fixation-like behaviour of the investigated VA^3V system by taking advantage from such evolutionary tuning. On the other hand, the advantage of the eDN and SalGAN scores against those measured for VA^3V system could partly be explained by the fact that in contrast to VA^3V system, these algorithms have been designed and exploited exactly as eye fixation estimation approaches, where such an estimation is the main goal.

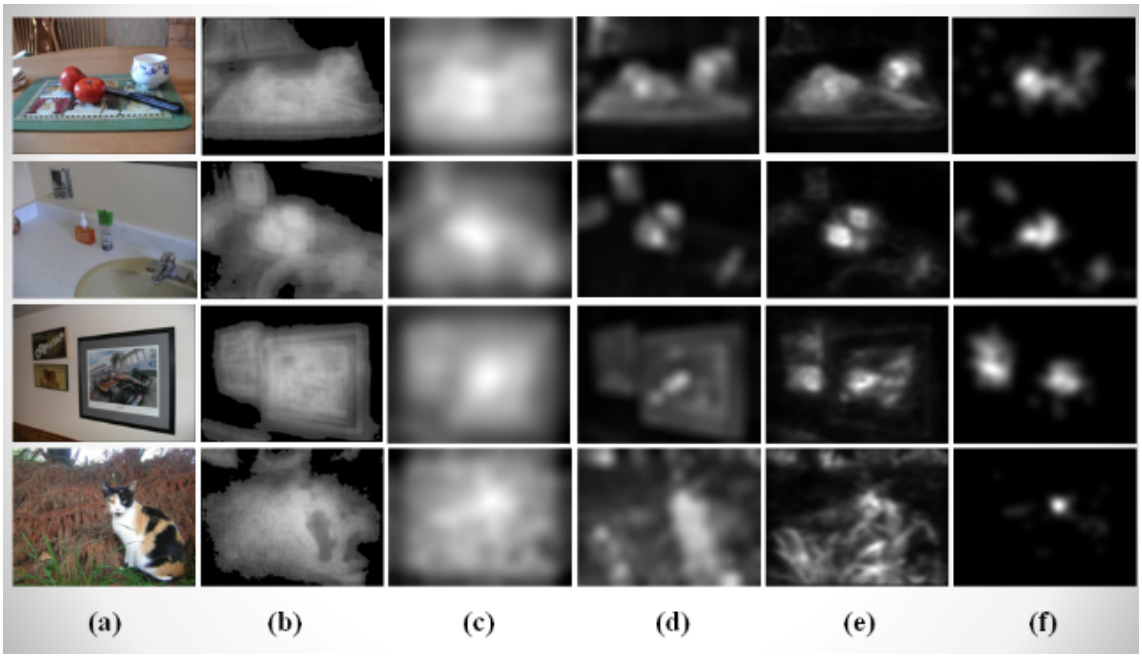

Figure 3.3: Examples of visual attention maps carried out by VA^3V system (b) and predicted eye-fixation maps achieved by eDN (c), BMS (d) and RARE2012 (e). Columns "a" and "f" give input stimulus and the corresponding grand true (e.g. experimental) eye-fixation maps, respectively

3.3.4 Generalization ability

Let us consider 60% of the MIT1003 benchmark database (e.g. 600 images selected randomly) as learning set which is used for tuning the system's parameters – this means that at each iteration of GA tuning we apply VA^3V model with the set of parameters of each chromosome to all and every image in the learning set, to be measured after by Fit integral evaluation indicator.

The learning phase is considered as completed, if the generation-best fitness function Fit of any chromosome either remains constant or decreases during several generations. The rest of the above-mentioned database's images (e.g. 403 images), joining together with those of TORONTO benchmark database (e.g. 120 images) have set up the testing dataset, collecting 523 20 images. The experiment has been repeated several times and the three best sets of tuned parameters have been retained: labeled "Run-1", "Run-2" and "Run-3", respectively.

The fitness function's evolution during the training (tuning) phase is illustrated by Figure 3.4, and Table 3.3 summarizes results both for training and for testing phases showing several parameters from "winning" chromosomes – WSC parameters, w_G weights, A and Nar parameters of Gaussian blobs, FT adaptive thresholds, – along with resulting assessment measures: fitness functions Fit and the usual in-

Table 3.3: Summary of the obtained results for tuning and testing phases

Run	Chromosome[1]	Dataset[2]	Fit	AUC_J	AUC_B	KL	NSS	$sAUC$
Run-1	(0.21, 0.22; 0.49; 245; 2.44; 38)	T-600	3.5653	0.8402	0.6549	0.6372	1.1747	0.5346
		V-523	N/A	0.8358	0.6429	0.5858	1.4126	0.5869
Run-2	(0.27; 0.49; 242; 1.77; 54)	T-600	3.5577	0.8312	0.6629	0.6675	1.1263	0.6074
		V-523	N/A	0.8319	0.6519	0.6147	1.2354	0.6038
Run-3	(0.18; 0.49; 213; 2.20; 50)	T-600	3.4074	0.8373	0.6486	0.6636	1.046	0.5414
		V-523	N/A	0.8342	0.6377	0.6107	1.3589	0.6569

[1] Each chromosome is represented as a set of parameters in the following order: (WSC; w_G; A; Nar; FT).
[2] "T-600" represents the tuning dataset, consisting of 600 MIT1003 dataset images, and "V-523" represents the testing (or validation) dataset, consisting of other $403 + 120$ images of MIT1003 and Toronto datasets.

dicators (AUC_J , AUC_B, KL, NSS and $sAUC$), respectively.

Confirming the concluding statements of the previously-described experimentations regarding the scores' magnitudes values (in term of their closeness to those of the state-of-art best algorithms' ones) and regarding the emergence of an eye fixation-like behavior of the investigated model, the first remark about this second experimental assessment regards validation of the evolutionary tuning process efficiency itself. However, it is essential to note the extent of the obtained scores in testing phase, being reminiscent that the involved dataset of images doesn't include any pattern (image) from training dataset.

This means that carried out visual attention maps keep a same degree of likeness to ground-true eye fixation maps that had been obtained for the learning dataset, and thus confirming the generalization ability of the investigated system.

3.3.5 GA-based Tuning Efficiency Assessment

For this, a subtle selection of 60 images from the MIT1003 benchmark database has been performed, creating a severely reduced new training dataset, containing 10% of the learning dataset used in previous experimentation. The selection criteria have been based on representativeness of the selected images regarding diversity of landscapes, colors' assortment and attractive objects' variety (e.g. objects which have focused in actual fact the humans' attention through the corresponding ground-true eye fixation maps). This new reduced dataset has been used for tuning the system's parameters. The learning phase is considered as completed if the highest fitness function either remains constant or decreases during three consecutive generations.

The examples of these 60 images are shown in Figure 3.5, where they are divided

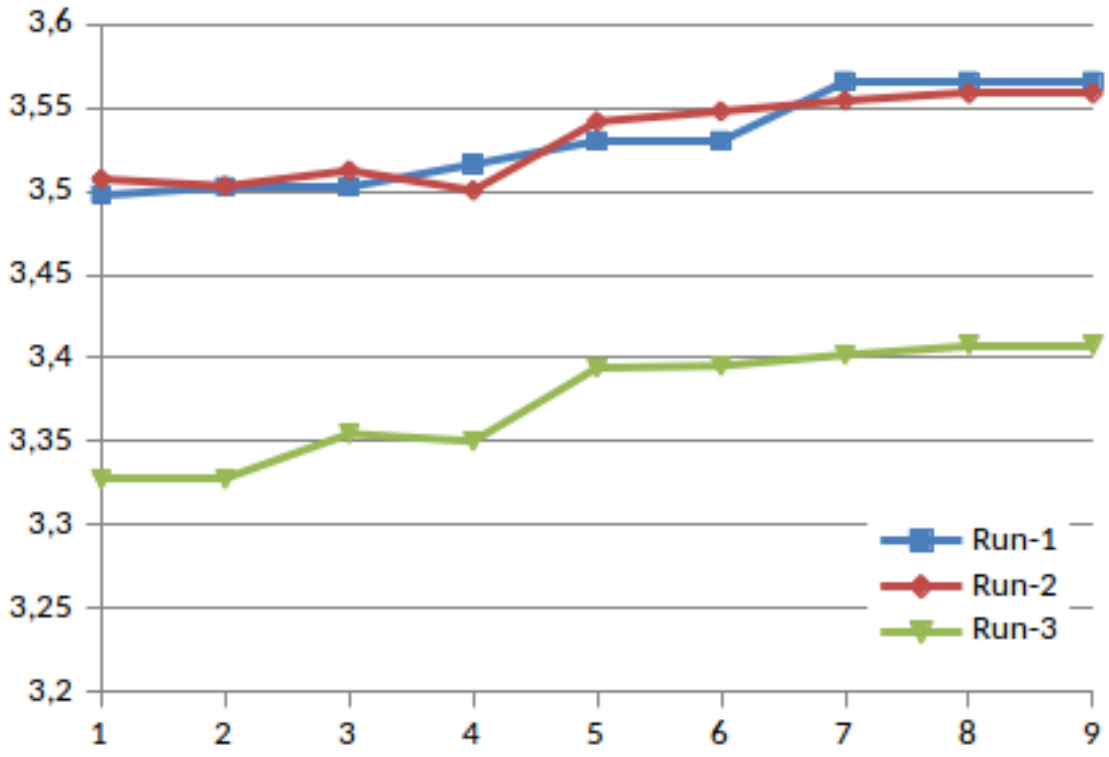

Figure 3.4: Fitness function's evolution during the tuning phase. The tuning dataset includes 600 images randomly selected from MIT1003 benchmark database

into four levels of subjective processing difficulty – with "Low" labelling the lowest level and "Very-High" the most difficult one, – and into several levels of image complexity – with "One" attractive object, with "Two", and with "More" than two. Also shown are corresponding ground true eye fixation maps, attractive objects and corresponding areas.

In the same way, a testing dataset, including also 60 images, has been built following the above-mentioned policy: i.e. including images with increasing "processing difficulty" and "image's complexity".

By "processing-difficulty", we mean the difficulty to predict the area (or object), which has mostly attracted human's eye fixation. For example, in the image representing a "red bird", the so-called red bird has been the unique attractive item (simple case with a low ambiguity concerning the attractive item). While concerning the image of the orange car, the attractive item has not been the "car" but its "license plate", so the processing difficulty depends on several non-formalized factors and is assessed by researchers in relative categories.

By "image's complexity", we mean the density of salient visual information of the image. In fact, an image containing only one salient object is considered simpler

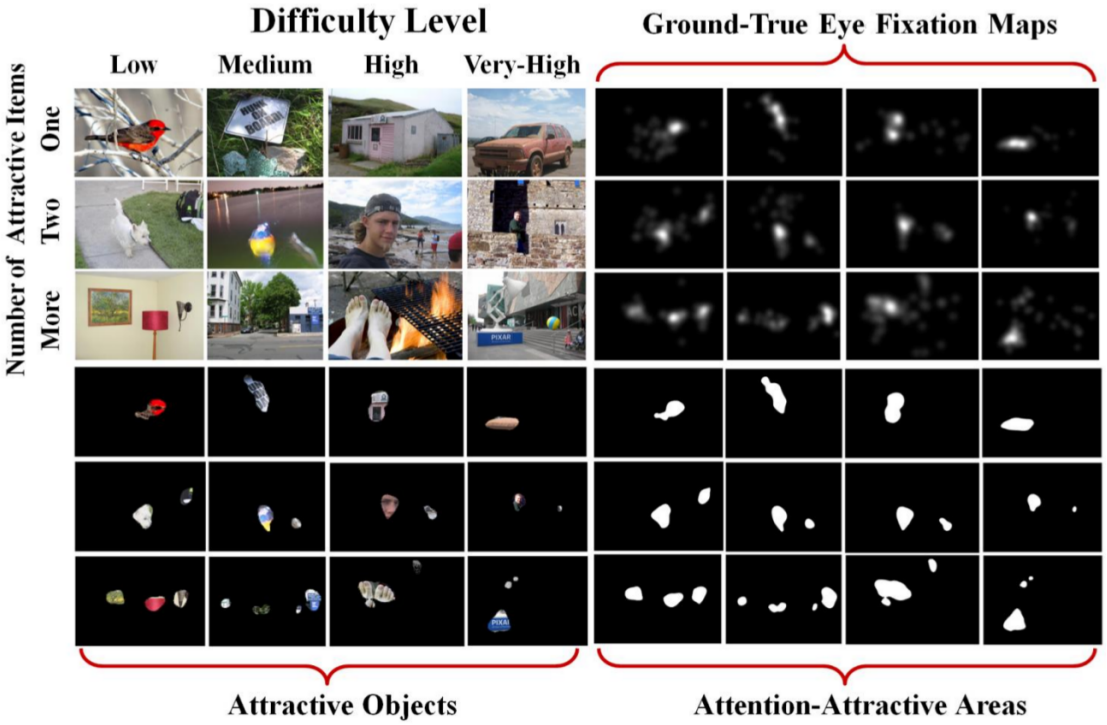

Figure 3.5: Examples of specifically selected images, representative of processing difficulty, of T-60 image subset of MIT1003 dataset

regarding the above-mentioned criterion that an image including several potentially salient objects which may attract (or not attract) the human's visual attention.

Not only the "cherry picked" testing dataset has been used; three testing datasets have been considered. The first one, containing the same amount of images as the learning dataset (e.g. also 60 images) has been constructed following the same policy used for building the learning dataset (e.g. a specific selection of 60 additional representative images from MIT1003 benchmark database). The TORONTO benchmark database has been considered as the second one.

Finally, the previously used testing database (e.g. the testing dataset collecting 523 images) has been considered as third testing dataset. In the same way, the three best sets of tuned parameters have been retained and labeled: "Run-4", "Run-5" and "Run-6", respectively. The Figure 3.6 illustrates the fitness function's evolution during the training (tuning) phase, including the three updating cycles, and Table 3.4 summarizes results as well for training as for testing phases, showing chromosome parameters values, fitness function and all other indicators.

Here also, regarding the closeness to state-of-art best algorithms' scores and regarding the emergence of an eye fixation-like behavior of the investigated model, the obtained scores confirm the concluding admissions of the previous experimen-

Table 3.4: Summary of the obtained results for tuning phase performed using the reduced learning dataset and testing results on three testing datasets.

Run	Chromosome	Dataset[1]	Fit	AUC_J	AUC_B	KL	NSS	$sAUC$
Run-4	(0.2, 0.36; 0.49; 246; 2.06; 89)	T-60	3.5829	0.8305	0.6833	0.7062	1.1612	0.6142
		V-60	N/A	0.7867	0.6460	0.6675	1.2415	0.6433
		V-120	N/A	0.7875	0.6236	0.4522	1.3471	0.6224
		V-523	N/A	0.8071	0.6536	0.5888	1.451	0.6211
Run-5	(0.17; 0.49; 167; 1.92; 60)	T-60	3.5687	0.8295	0.6814	0.7072	1.1874	0.5916
		V-60	N/A	0.8089	0.6437	0.6438	1.1967	0.6182
		V-120	N/A	0.8144	0.6178	0.4608	1.0282	0.6055
		V-523	N/A	0.8205	0.6481	0.6005	1.2041	0.5976
Run-6	(0.2, 0.33; 0.45; 246; 2.06; 73)	T-60	3.4929	0.8409	0.6803	0.7267	1.0923	0.6281
		V-60	N/A	0.8122	0.6450	0.6529	1.238	0.6308
		V-120	N/A	0.8304	0.6235	0.4459	1.3188	0.6364
		V-523	N/A	0.8307	0.6520	0.5904	1.3218	0.6732

[1] "T-60" represents the 60-image tuning dataset; "V-60" represents the 60-image testing (or validation) dataset;
"V-120" represents the TORONTO dataset, used for validation; "V-523" represents the 403+120 images of MIT1003 and Toronto datasets.

tations. We should note the extent of the obtained scores in testing phase, being reminiscent that the involved testing dataset has drastically been reduced, including only 60 images, even though specifically selected. Thus, additionally to validating once more the outstanding generalization ability of VA^3V model, these results show also quite excellent efficiency of the incorporated GA-based tuning strategy in spite of the usage of the aforementioned considerably poorer tuning dataset.

3.3.6 Best Predictions as a Quasi-Saccade Model

As we call the map, resulting from VA^3V model calculation, a VAM (visual attention map), and argue that an adequate tuning of the panel of the involved parameters may shove the behavior of the generic eye fixation mechanism towards a "human-like" gazing behavior, – one may also interpret the regions corresponding to highest probabilities (i.e. highest values) of the resulted VAM as acting for artificial eye fixation areas simulating a human-like artificial gazing. To extend this interpretation, Let us define a "high-probability point" as a point, located at position HPP_i within VAM, for which the value $M_{\text{VAM}}(HPP_i)$ of the corresponding pixel in the map is not lower, than at least $N\%$ of the highest value of the whole VAM, as expressed in Equation 3.23, where $\widehat{HPP}$ represents a subset of Ψ^2 image pixels ($\widehat{HPP} \subset \Psi^2$) with each pixel being a "high probability pixel".

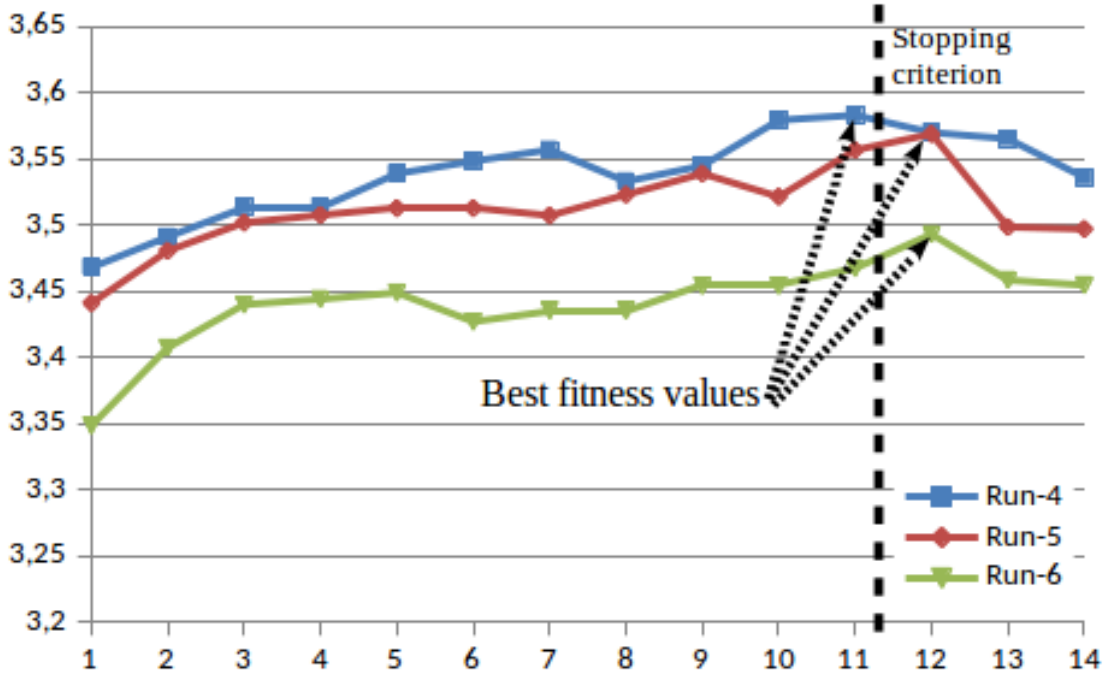

Figure 3.6: Fitness function's evolution during the tuning phase using the reduced learning dataset, including 60 specifically selected images from MIT1003 benchmark database.

$$\forall X \in \Psi^2, X \in \widehat{HPP} \text{ if } \mathrm{M_{VAM}(X)} \geq \frac{\mathrm{N}}{100} \max_{\mathrm{K} \in \Psi^2} \mathrm{M_{VAM}(K)} \tag{3.23}$$

Let us define a "high-probability region" (denoted as $HPR_k \subseteq \widehat{HPP}$): a semi-linked area of high probability points, where each high probability point HPP_i, belonging to HPR_k, has at least one neighbour HPP_j (belonging to the same region), which is located within a neighborhood characterized by an Euclidean distance lower, than a threshold, as shown in Equation 3.24.

$$d(HPP_i, HPP_j) \leq d^*, \quad \forall HPP_i \in HPR_k, \quad \exists HPP_j \in HPR_k \, (i \neq j) \tag{3.24}$$

Then for each region HPR_j there should exist at least one point with highest pixel intensity value, which we could call a "representative point of j-th region", which we denote as $R_j \in HPR_j$, as defined in Equation 3.25. This point admits the highest human gaze catch predicted probability over all points of this region.

$$M_{\mathrm{VAM}}(R_j) \geq M_{\mathrm{VAM}}(HPP_i), \quad \forall HPP_i \in HPR_j \tag{3.25}$$

In other words, all high probability points, located at an Euclidean distance farther, than d^* from R_j, will belong to another region HPR_k. In this case an iterative algorithm of finding all the "representative" points of the map can be defined, which in general would be very similar to an iterative algorithm of finding local maximas of a function defined on two variables.

Algorithm 1 "Representative" points search

1: $MaxIntensity \leftarrow \max_{K \in \Psi^2} M_{\mathrm{VAM}}(K)$
2: $X \leftarrow$ all pixels, where $M_{\mathrm{VAM}}(X) \geq \frac{N \times MaxIntensity}{100}$
3: $OrderedX \leftarrow \mathrm{SortByIntensityDesc}(X)$
4: $RepPoints \leftarrow \emptyset$
5: $OtherPoints \leftarrow \emptyset$
6: **for each** $x \in OrderedX$ **do**
7: **if** $x \in OtherPoints$ **then continue**
8: $RepPoints \leftarrow x$
9: $OtherPoints \leftarrow$ any y where $d(x,\ y) \leq d^*$

If Euclidean distance and its threshold d^* are measured in terms of number of pixels, then it could be expressed regarding the image size $n \times m$ in pixels as $d^* = C_{\mathrm{HPP}}\sqrt{n \times m}$, where C_{HPP} represents a tunable coefficient invariant of image size.

According to this extension of a human-like vision model interpretation, we could argue that the representative points, showing the "best prediction" points, may imitate some kind of saccadic fixation positions of the human eye, as these points represent not only most salient pixels, but the salient regions – if an eye stops its gaze on such a point, than due to its central-periferic antagonism it could catch the small region around this point and than make another saccade to another "representative" point in another part of the image.

Continuing with such an interpretation, assuming that any Hamiltonian path in a complete graph, drawn over all representative points in VAM can estimate a saccadic dynamic (i.e. movements) of a generic eye fixation behavior, we argue that within the above-mentioned frame, the appropriate tuning of the involved parameters will shove the model's behavior toward human-like eye fixation behavior, modeling the humans' way of gazing the surrounding landscape. In other words, the presented model could approach the eye fixation dynamics of human's visual gazing. In fact, the human's gazing can produce up to 50 saccadic movements (i.e. eye fixations) per second, while mostly averaging to 4-8 per second [Hamm 10].

3.4 Combining the Approaches

3.4.1 Top-Down Recognition-based Model

As we have already discussed a theoretical basis of the top-down attention model in section 2.4, and the informed choice of approaches to be used has already been outlined in subsection 1.4.1, this subsection considers only two other possible recognition approaches – the fine keypoint-based recognition, and pattern-based recognition.

3.4.1.1 Fine-grain Recognition: Confidence Levels

The keypoint-based algorithms are not formally designed to produce any confidence level; yet, the matching between given inputs and stored images is based on Euclidean distance between descriptor vectors (for decimal-based descriptor algorithms, as SURF) or Hamming distance between descriptor vectors (for binary-based descriptor algorithms, as BRISK).

Also, there is a complimentary matching algorithm, which produces the decision – whether keypoints with descriptors, found in input image, really represent the image from memory; it is a de-facto standard in the computer vision domain to apply KNN-based matcher [Lowe 04]. It applies to all the descriptor vectors' distances, and (under additional constraints) if the best match gives distance lower, than a fraction of next best distance, then this match is found to be true. A scheme in Figure 3.7 depicts the order of work for this module. Here input image is processed by the descriptor extraction algorithm (SURF or BRISK), as well as each existing image in the "library" (in our case this is a storage of already learnt objects, based on the visual sketchpad of semantic network). Then the sets of descriptors are matched between themselves; if any set from library is classified as "matching" with any input subset of descriptors, this subset is interpreted as "recognized".

While in subsubsection 1.4.1.1 we stated, that efficiency-wise the BRISK algorithm should be used, we should also pose the question of implementation. Existing open source implementations of BRISK and SURF (OpenCV library, [Bradski 00]) give an exhausting answer: licensed SURF implementation produces descriptors much faster, than BRISK. In our experiments on 8-core CPU SURF implementation used multithreaded calculation, while BRISK implementation used only single CPU, resulting with SURF having almost two times faster computation of descriptors than BRISK for the same set 50 images from "Things-50" database with almost any set of parameters. These results provide us an ambiguity in interpretation, jus-

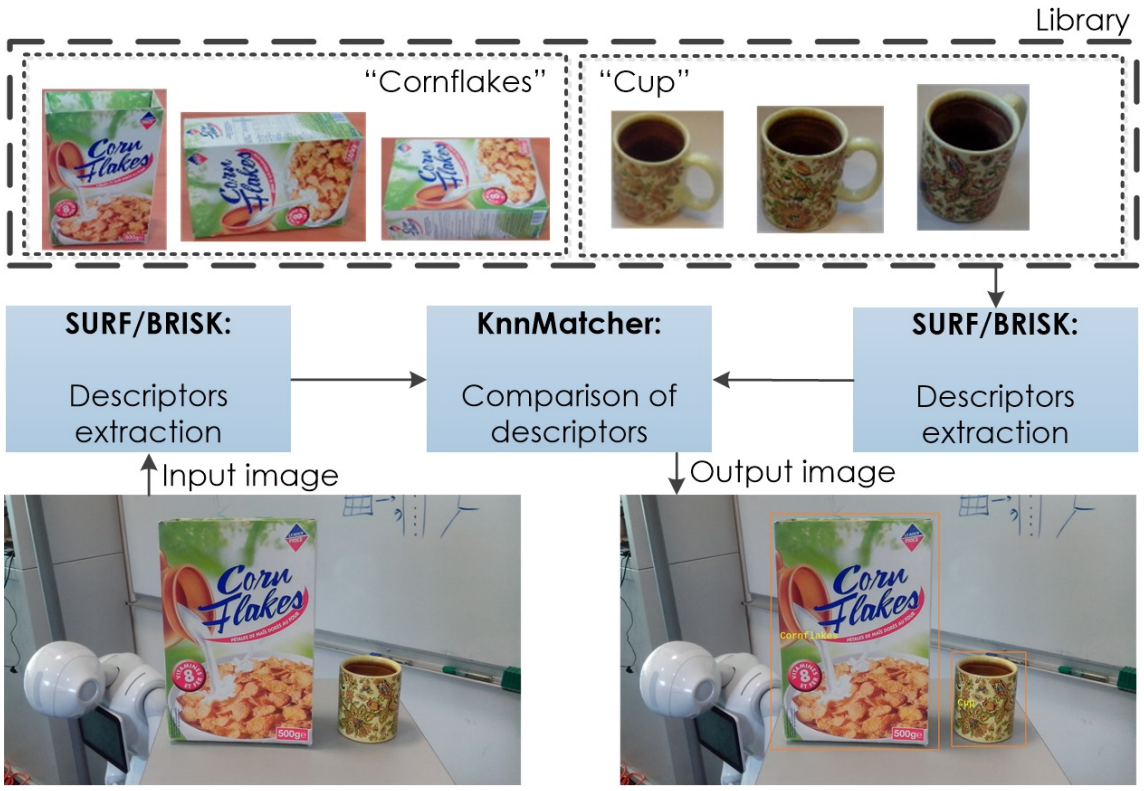

Figure 3.7: An outline of the keypoint-based recognition module

tifying as well usage of SURF if we plan to have faster implementation.

For the rest of this work we will stick with our first choice, BRISK, based on the idea, that better and faster implementations of this algorithm could arrive in future.

The matching algorithm is based on initial notions from [Lowe 04], as well as [Ozuysal 10], and some parameters previously used by [Ramík 12] and [Hassan 15]. The KNN-based matcher takes each descriptor from input, produces Euclidean or Hamming distances for this descriptor and all library image descriptors; if the relative distance of best-matching pair is lower, than the distance of the second-best matching pair multiplied by threshold ratio D, this pair is considered as "probably matching". After this stage, if there is at least N "probably matching" pairs of descriptors, the corresponding keypoints from input image are compared against the keypoints of the library image in order to find homology with RANSAC algorithm [Chum 03]. If the homology exists, the object is considered as recognized.

As the distances between descriptors are multiple relative values, we can not use them to derive the confidence level of recognition; the only real confidence-relative characteristic is the number of matched keypoints. The lowest level of matched keypoints for the recognized object is N, the highest level is 128 (the total number of keypoints, produced by both descriptor extraction algorithms by default). Also the

confidence level depends on the ratio threshold D, because the higher is threshold, the lower should be the confidence. We made several simulations with "Things-50" database in order to find the highest-achieved absolute numbers in both true positive and false positive classifications, depending on the thresholds N and $D\%$; if treating TPR as a confidence level proxy, we can derive an equation for it, as shown in Equation 3.26, by extrapolating it as Richards' curve. Here KP represents the real number of matched keypoints.

$$CL_{BRISK} = \frac{1}{\frac{D}{1-D} + e^{-D(KP-N)}} \tag{3.26}$$

3.4.1.2 Pattern-based Recognition: Parameters to Use

The Viola-Jones framework acts as a very open platform for experimenting with parameters. A simple GA-based training on the **Faces-400**, – dataset, which is a subset of MIT1003 dataset, containing faces, – with fitness function based on Precision/Recall in Wolf interpretation, gives us pretty good results as shown in Table 3.5. Here parameter sets are shown in form of (Scale; Minimal region size; Minimal Neighbours), where scale depicts the resizing allowance for the algorithm, minimal region size is given in form of a fraction of the initial image size (e.g., $''0.05 \times 0.05''$ for 1024×768 image is equal to 51×38), and minimal neighbours value stands for the level of how many neighbouring true detections in the "almost" same region should be found in order to declare this region as true detection.

Table 3.5: Example sets of parameters of Viola-Jones framework, and their performance on Faces-400 dataset

Parameters	Pr_{Wolf}	Rec_{Wolf}
1.2-0.05x0.05-4	0.86	0.53
1.2-0.08x0.08-6	0.715	0.648
1.2-0.1x0.1-8	0.632	0.71
After GA Training [1]	0.851	0.62

[1] The best GA-trained results are found to be (1.4-0.06x0.06-4)

Also Figure 3.8 shows several examples from our quasi-real-world simulations, as how the recognition results are depicted.

Figure 3.8: Examples of recognition in quasi-real-world simulations: previously known objects and human faces

3.4.2 Short-Term Memory and Decision Model

In section 2.5 we have already outlined a schema for memory and decision module, based on Baddeley-Hitch schema. This section contributes to enhancement of this analogy by using some existing meta-heuristics for each part of this schema.

As subsection 1.4.2 has already depicted the usability of Wu-Palmer distance for lexical comparison between different (similar) concepts, there is not much more to discuss in terms of implementation. While it is an important part, it is more an auxiliary unit for the central executive – as it provides the quantitative means for comparison of otherly non-comparable concepts; this means, that "Phonological Loop" can be presented as WordNet usage in terms of this simple semantic analysis.

3.4.2.1 Visual Sketchpad: knowledge storage

Let us investigate the usability of models and mechanisms, used in the field of knowledge representation, in the context of this work. While we remember, that STM conducts bidirectional connection with TD unit and its recognition part, following premises can be found:

- Information, extracted from visual perception, will definitely be represented in the form of an image or set of images, along with additional meta-information, derived by recognition module;
- new knowledge, created autonomously "without teacher" (e.g., in recognition module), can not be treated as 100% true – due to inexistence of a model, which can transform the visual data into textual with 100% efficiency;
- an existing system of the knowledge, which are already gained, should have the most general form, – due to the lack of initial context of tasks, which should be done by the robot; inherently, one should assume a general context, until the opposite is given.

According to the given premises, we can hypothesize about usability of a semantic network, with a small set of used general relations, as a basic construct of data and knowledge storage working as a "visual sketchpad". Upon such memory, more specialized mechanisms could be applied.

The creation of a simple semantic memory, which would supply the data to the fine-grain recognition algorithm, could be based on several assumptions:

1. An object can be learnt and remembered, if it is found by a segmentation algorithm, and recognized by broad recognition CNN with confidence level of one of the recognition labels L higher, than a threshold CT ($CL(L) \geq CT$);

2. If there is no object/concept with label L in the memory, – add it there, establishing a relation "possibly seen with" with all the objects/concepts, found in the same image, with weight 1. Remember also the confidence level;

3. If there exists an object/concept with such label in the memory, – add the newly acquired image to the memory for the same object/concept, interpreting this as another possible outer view of this object/concept (also updating the relations – for all unmet relations the weight is penalized by coefficient of WPT).

The segmentation algorithm, mentioned in first assumption, can be chosen arbitrarily – e.g., Comaniciu's Mean-Shift [Comaniciu 02] or Moreño's Spherical Coordinates-based algorithm [Moreno 11].

An example of simple semantic network in shown in Figure 3.9 – where an image is processed to achieve three new nodes in the network, connected by relations of "possibly seen with".

3.4.2.2 Episodic Buffer: Growing Self-Organizing Maps

As the episodic buffer is not strictly defined to address an exact problem (this is the domain of Central Executive), we can define this part of the memory model as the one which creates a gist-like information about each input image in the overall sequence, while learning them and finding hidden similarities in order to give additional info to the Central Executive about the change of environment.

An ideal candidate for such a work is a Growing Self-Organized Map (also known as Kohonen map), which can produce gist-like comparison of sequence images on the fly. Thus it can be justified as a mechanism of episodic buffer which is able to answer the question *"Is the situation around the robot similar to the one of the*

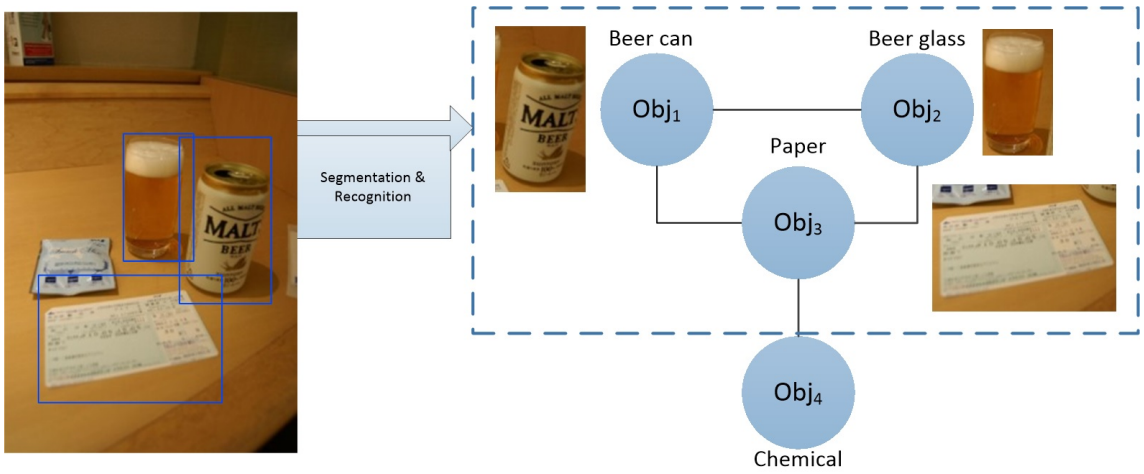

Figure 3.9: An example of a simple semantic network

previous situations?", and by this impose the status-quo where the visual attention mechanism general results should be comparable to those, acquired previously.

A Self-Organized Map (SOM) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional discretized representation of the input space of the training samples, and is therefore a method to do dimensionality reduction.

It has been used countless times in many different domains, being one of the soft-computing meta-heuristics de facto. E.g., it has been used in image processing for segmentation [Haring 94] or image compression [Khan 14]; yet, applying similar conditions, we try to use as a clusterisation tool over the image input.

A growing self-organizing map (GSOM) is a growing version of SOM. The GSOM was developed to address the issue of identifying a suitable map size in the SOM. It starts with a minimal number of nodes (usually 4) and grows new nodes on the boundary based on a heuristic. By using the value called Spread Factor (SF), the data analyst has the ability to control the growth of the GSOM.

All the starting nodes of the GSOM are boundary nodes, i.e. each node has the freedom to grow in its own direction at the beginning. New Nodes are grown from the boundary nodes. Once a node is selected for growing all its free neighbouring positions will be grown new nodes. In GSOM, input vectors are organized into categories depending on their similarity to each other. For information reduction, the image or data is broken down into smaller vectors for use as input. For each input vector presented, the Euclidean distance to all the output nodes are computed. The weights of the node with the minimum distance, along with its neighbouring nodes are adjusted. This ensures that the output of these nodes is slightly enhanced.

This process is repeated until some criterion for termination is reached.

In general, the functioning of a GSOM can be outlined as follows (as adopted from [Alahakoon 00]):

- **Initialisation phase:** Initialize the weight vectors of the starting nodes (usually four) with random numbers between 0 and 1; calculate the growth threshold (GT) for the given data set of dimension D according to the spread factor SF using the formula $GT = -D \times \ln SF$);
- **Growing phase:** Start of iteration: present the input data to the network.
  1. Determine the weight vector that is closest to the input vector mapped to the current feature map (winner), using Euclidean distance (step, similar to the SOM). This step can be summarized as: find q' such that $|v - w_{q'}| \leq |v - w_q| \, \forall q \in \mathbb{N}$ where v, w are the input and weight vectors respectively, q is the position vector for nodes and N is the set of natural numbers.
  2. The weight vector adaptation is applied only to the neighbourhood of the winner and the winner itself. The neighbourhood is a set of neurons around the winner, but in the GSOM the starting neighbourhood selected for weight adaptation is smaller compared to the SOM (localized weight adaptation). The amount of adaptation (learning rate) is also reduced exponentially over the iterations. Even within the neighbourhood, weights that are closer to the winner are adapted more than those further away. The weight adaptation can be described by Equation 3.27, where the Learning Rate $LR(k)$, $k \in \mathbb{N}$ is a sequence of positive parameters converging to zero as $k \to \infty$. $w_j(k)$, $w_j(k+1)$ are the weight vectors of the node j before and after the adaptation, and N_{k+1} is the neighbourhood of the winning neuron at the $(k+1)$-th iteration.

$$w_j(k+1) = \begin{cases} w_j(k) & \text{if } j \notin \mathrm{N}_{k+1} \\ w_j(k) + LR(k) \times (x_k - w_j(k)) & \text{if } j \in \mathrm{N}_{k+1} \end{cases} \tag{3.27}$$

     The decreasing value of $LR(k)$ in the GSOM depends on the number of nodes existing in the map at time k.
  3. Increase the error value TE_j of the winning neuron j (error value is the difference between the input vector and the weight vectors).

4. When $TE_i > GT$ (where TE_i is the total error of node i and GT is the growth threshold), grow nodes if i is a boundary node, or distribute weights to neighbours if i is a non-boundary node.
5. Initialize the new node weight vectors to match the neighbouring node weights, initialize the learning rate LR to its starting value.
6. Repeat steps 1–5 until all inputs have been presented and node growth is reduced to a minimum level.

- **Smoothing phase:** Reduce learning rate and fix a small starting neighbourhood; find winner and adapt the weights of the winner and neighbours in the same way as in growing phase.

In order to use GSOM for image processing, we need to establish both the architecture used, as well as several parameters.

Assuming we take the input image of $n \times m$ size given in RGB format, we need to transform it into a vector of real values, usable as an input for SOM. One of the most simple approaches, usable for this, is to resize the image by several times, divide the resulting image into $n' \times m'$ tiles, transform it into YCC color space and find the average value of luminance intensity for each tile. These intensity values, after normalization, if interpreted as a 1-dimensional vector, can be used as input vector for SOM where SOM has exactly $n' \times m'$ neurons at the input layer.

Another approach suggests the usage of additional algorithms for best information reduction: principal component analysis, Gabor filtering, etc. We apply the so-called "perceptual hash" approach [Zauner 10], which implies that the resulting hash for similar images should be similar. This approach generally speaking represents almost the same idea of reducing the information positionally; Figure 3.10 represents several comparisons of perceptual hash.

Usually the comparison between perceptual hashes is done as calculation of Hamming distances, giving distance 3 for the left column, 10 for the center and 24 for the right column. On the other hand, a threshold for the decision whether the images are similar or not, is not provided and can be parametrized by each user of the algorithm.

Our threshold here is the usage of GSOM with the perceptual hash input (neighbourhood radius = 3, number of input nodes = 128, maximum number of output nodes = 100); the results on the simulation data correlate with mean square error comparison metric.

Figure 3.10: Some images from our datasets with corresponding perceptual hashes. Left column shows two almost identical images, center column represents two images with one different region in center, and right column provides two images which show different environment yet in the same room

3.4.2.3 Central Executive: Decision Making

In order to describe a decision making module, we need to introduce a quantitative measure which will be able to help us in this case.

Assume there exists a quantity of importance I_{img}, given to the whole image, as well as importance I_j for each object j, found (or probably found) on this image. Comparison of importance between objects, found in the same image, can outline difference in relative saliency for these objects, and thus let the decision module apply this information to the resulting mix of attention maps.

Importance of different types of objects definitely changes for different tasks, thus quasi-applying third sub-direction of "top-down" visual attention. Assuming the tasks:

- Generic task: *"free visual search"*, look for humans or texts in order to change the task.

$$I(''HumanFace'') = 1,\ I(''Text'') = 1,\ I(Other) = 0 \tag{3.28}$$

- Search task: *"look for a specific object (or, maybe, something similar)"*, as shown in Equation 3.29.

$$I(Obj) = \begin{cases} 1 & \text{if } Obj \in \left\{SpecObj\right\} \\ Imp * wup(SpecObj, Obj) * CL(Obj) & \text{if } Obj \notin \left\{SpecObj\right\} \end{cases} \tag{3.29}$$

Here $wup(SpecObj, Obj)$ represents Wu-Palmer similarity of objects over WordNet ontology ("Language analysis"), and $CL(Obj)$ represents confidence level of recognition, as given by the recognition module. Imp can be defined as an importance decrease coefficient and be applied if needed. Default value for it should be $Imp = 1$, if we allow other objects to be seen as possibly important. If we imply strict conditions of search, then Imp should be 0.

- Social task: *"Interact with humans"*, as an inter-modification of generic and search tasks.

$$I(Obj) = \begin{cases} 1 & \text{if } Obj \in \left\{''HumanFace''\right\} \\ Imp * wup(SpecObj, Obj) * CL(Obj) & \text{if } Obj \notin \left\{''HumanFace''\right\} \end{cases} \tag{3.30}$$

Here we assume that there is a possibility, if a part of human body is recognized (e.g., elbow), there might be human face nearby, therefore such an object might be important.

Another point of view is the overall image's importance level $I_{img}(k)$, given by episodic buffer: whether the k-th given input image represents high level of similarity to $k-1$-th input image, the importance of this image should diminish: the robot has already seen all this, and there probably is nothing new in the visual field. In this case any change of visual scene would provide surge of image's importance; assuming j_k as winning neuron at iteration k, and j_{k-1} as winning neuron at iteration $k-1$, the importance of input image $I(k)$ is as given by Equation 3.31.

$$I_{img}(k) = \begin{cases} 1 & \text{if } j_k \notin \mathrm{N}_{k-1} \text{ or } (k = 0) \\ Imp_K * I_{img}(k-1) & \text{if } j \in \mathrm{N}_{k-1} \end{cases} \tag{3.31}$$

Here Imp_K represents diminishing coefficient of "curiosity drop", – the lower it is, the faster will diminish the importance of similar environment.

Thus we can formulate a top-down visual attention map ($M_{TD}(k)$) for image k, as given in Equation 3.32, where $M_{TDI}(k)$ represents "map of importance" – a heatmap, close to saliency map in its idea (as already defined in the scope of this

work in section 1.3), and defined for each pixel x as importance level of the object $Obj(x)$, if the pixel x is a part of representation of the object $Obj(x)$. If there is no object found around pixel x, than $Obj(x) = \emptyset$, and $I(Obj(x)) = I(\emptyset) = 0$.

$$M_{TDI}(k, x) = I(Obj(x)) \tag{3.32}$$

$$M_{TD}(k) = I_{img}(k) * M_{TDI}(k) \tag{3.33}$$

Thus, in the output of the whole top-down part of system, processing the k-th image, there will be:

- Map of importance, $M_{TD}(k)$
- Set of n objects, found at this image, $\left\{Obj_{k1}, \ldots, Obj_{kn}\right\}$
- Set of confidence levels of recognition of these objects, $\left\{CL(Obj_{k1}), \ldots, CL(Obj_{kn})\right\}$
- Overall importance estimation of this image, $I(k)$

Here only first part, map of importance, will be used by default by the visual attention model in order to produce the combined visual attention map as a fusion of bottom-up saliency map and top-down importance map.

Other parts can be given to other reasoning modules of bigger robot actuating scheme, in order for it to react.

3.4.3 Assembling the Modules Together

Let us redefine the more detailed scheme of the whole system, as shown in Figure 3.11.

To wrap up the theoretical discussion of the whole combined model, we need to outline the algorithm of its work:

1. Input image k is given to the bottom-up module (VA^3V model), which produces low-level attention map $M_{\text{VAM}}(k)$, along with the map of representative points.
2. The task and semantic memory are analysed by WordNet & Central Executive: whether there are specific types of objects, for which the importance would be non-zero? If so, which objects, already existing in memory, would produce such importance?

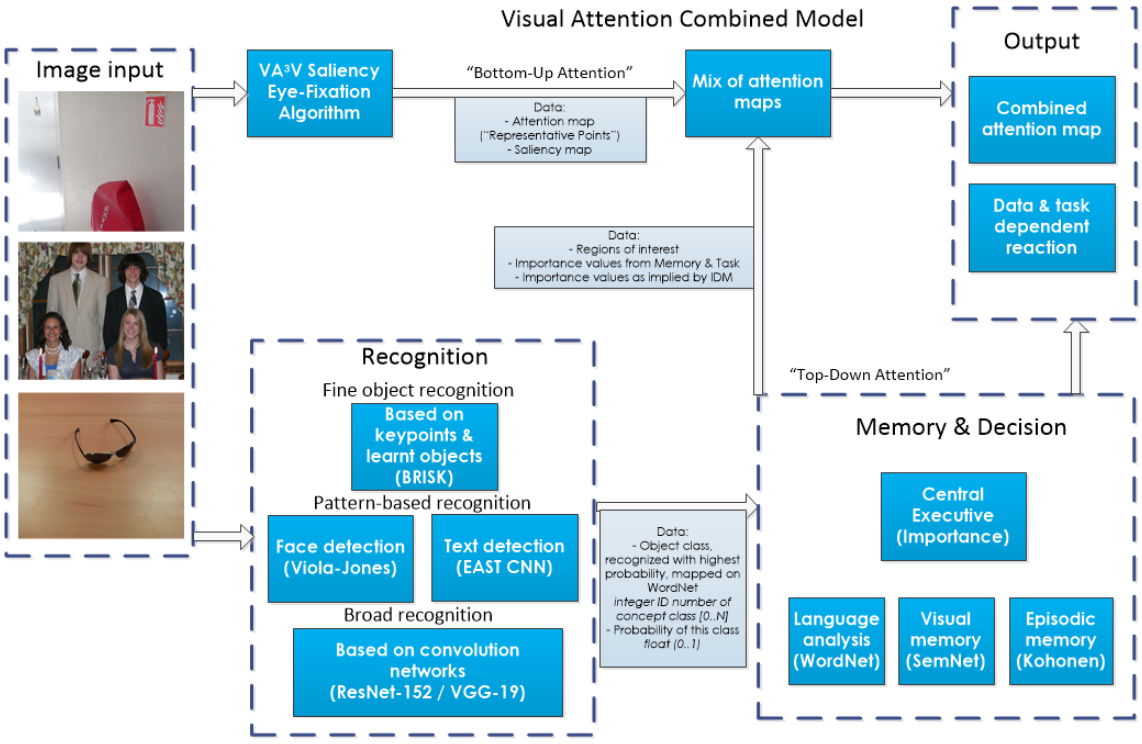

Figure 3.11: An extended human-like visual attention model, combining both attention directions: detailed vision, based on techniques choices

3. Input image k is given to the recognition module, where BRISK algorithm tries to find the possibly-important objects in it. For each of n_1 found objects – produce confidence levels, $CL_{BRISK}(Obj_n)$, $n \in (1, ..n_1)$

4. Pattern-based algorithms search for special objects – human faces and texts. For each of n_2 found pattern-based objects – produce confidence levels, $CL_{PAT}(Obj_n)$, $n \in (1, ..n_2)$

5. If steps 3 and 4 did not give any result ($n_1 = n_2 = 0$), apply broad recognition in order to produce top-5 assumptions about visual scene ($n_3 = 5$, and the confidence levels are given by the CNN algorithm itself $CL_{CNN}(Obj_n)$, $n \in (1, ..n_3)$).

6. If confidence level, produced in step 5, is lower than a threshold CT ($CT > \max_{n \in (1..n_3)}(CL_{CNN}(Obj_n))$), apply an image segmentation algorithm and try to recognize the biggest object found via step 5 (input changes from the whole image to the salient region with biggest area; the algorithms of detecting salient regions can be found in [Ramík 12]).

7. Input image is given to the memory module, where it is analysed by episodic

memory (GSOM), which produces image importance estimation $I_{img}(k)$.

8. If $I_{img}(k) = 1$, and steps 5–6 produced any object with a label, which confidence level CL was higher than CT, add this object to the visual memory.

9. Produce $M_{TD}(k)$ by algorithm, given in subsubsection 3.4.2.3, as an output of TD and decision making modules.

10. Produce the final, combined visual attention map $M_(k)$, based on following (where J represents all-ones matrix of the same size, as $I_(k)$, and $\odot$ represents the Hadamard product, or element-wise product of the matrices):

$$M(k) = I(k) \odot M_{TD}(k) + (J - I(k)) \odot M_{VAM}(k) \tag{3.34}$$

3.5 Conclusion

While the presented approach was described in Chapter 2 in general, this chapter is dedicated to detailed view on its elements, such as:

- A low level visual attention model is proposed, benefiting from the visual saliency as an implementation of several psychoneurological concepts. It has the capacity of human gaze prediction on the level, comparable to the results of state-of-art algorithms, yet being less time complex ($O(N * \log(N) + N) \leq O(M * P * I + P * I \log I + M * I * \log I)$, which implies at least 50% higher speed of processing if correctly implemented). It has been inspired by existing works studying the way human visually perceives his surroundings. In this context an approach, which is based on previous works and extends them, is suggested. This approach takes advantage of using the previously observed phenomena in human neurology, such as centred bias or center-surround antagonism. The algorithm has low complexity and can be run in real-time on contemporary processors, if implemented correctly. Moreover, it exhibits robustness to difficult real-world light conditions due to its inheritance from previous algorithm, invariant to light changes;

- The aforementioned low level model also shows a certain degree of flexibility by being able to be tuned for different input image types via genetic algorithm, also representing ability for generalization, which have been verified by quantitative evaluation over subsets of benchmark datasets MIT1003 and

TORONTO. The results show the quality of eye fixation prediction comparable to the best state-of-art algorithms, which allows us to extend the "human-likeness" interpretation into estimation of quasi-saccades;

- The presented model is the key part of the proposed Visual Attention Combined Model, which is also detailed in this chapter. Its results can be treated both as solutions themselves in the field of eye fixation problem (in any and each of its interpretations), and as one of the steps in modelling human visual attention on a higher level;
- The rest of the combined approach is described as a set of object recognition techniques, taken in ensemble for improved efficiency, as well as the moderating Decision And Memory Module which explores the means of the combination of approaches.

In future the low-level approach could evolve in several ways. As there exists a trend into the third, "neural network based" type of visual saliency models, one day (mostly in long-term perspective) the proposed approach might be heavily modified with the NN-based techniques (such as presented in [Liu 16] or [Pan 17]), if there is a possibility to run such an algorithm on the embedded systems in real time.

4 | Validation and applications

4.1 Introduction

Chapters 2 and 3 present a theoretical basis of this research, showing several aspects of a combined visual attention model, and approaches, usable as modules in such a system. But the algorithms presented – either ours, or third-party given, – are vastly different, and present themselves as standalone parts, it is inevitable to make them work together in order to show a proof-of-concept implementation of the whole model, working as a client-server system "mobile robotic platform"–"remote high-performance computer", as well as valorize and validate this model with several real-world experiments, implementing some semi-autonomous behaviours based on the visual attention.

Chapter 4 familiarizes the reader with the robots, used in the experiments, as well as shows general information about our implementation of the model, with some in-depth notes about several details concerning efficiency.

First part of the chapter, section 4.2, concerns general implementation of the model, as well as some details concerning efficiency. Next part, section 4.3, concerns a context of visual fire detection as a problem context, which is well-suited for autonomous BU validation, both in simulation and in real world experimentations. The section 4.4 is dedicated to several other real-world experiments focused on validation of the whole combined attention model and its usability. The chapter is then concluded with section 4.5.

4.2 Notes on System Implementation

As the priority is to explore usability of VA^3V bottom-up vision model in concern of robotic applications and real-time (or quasi-real-time) processing on both the embedded CPUs and remote processing CPUs, this gives us a notion that we need to explore, or at least describe, time complexity of this vision model and several practical advices for its usage. Another point is the difference of the robotic platforms used

(as we will discuss more thoroughly in subsection 4.3.1 and subsection 4.4.1): while Wifibot-M and NAO possess their own simple CPUs, their performance is seriously outdated and not capable of running such a model in real-time, thus demanding a remote computation architecture. On the other hand, Pepper robot's CPU is much more faster and, if using some pre-conditions, can run the "bottom-up" part of the extended model in quasi-real-time, while still asking for remote architecture in the "top-down" part.

Here we try to show an approach to evaluate theoretical time efficiency of the model's implementation and the possible effect of practical advices.

The implementation of the investigated system has been done as a set of cross-platform modules developed on Python language (version 2.7) and tested both on Linux (Ubuntu 16.04 LTS) and Windows (version 10) platforms (on remote computer), along with the robot Pepper embedded system (Gentoo Linux/OpenNAO distribution). Following subsections describe the choices of techniques, made in implementation, along with some time efficiency analysis.

4.2.1 Scaling the Input

Any notes on implementation of any algorithm cannot be full without notes about its time efficiency. Current experimental implementation is done in Python (with usage of Cython [Behnel 11]), and its time consumption is measured on Intel Atom 1.6 GHz (non-parallel due to Python restrictions) + DDR3-1600 4 Gb RAM system. It is worth noting, that further propagation of time efficiency is possible via added parallelization and/or full code translation into faster languages like C.

The primary impact on time consumption is given by the size of processed image, secondary – by used parameters in different parts of the model. We can estimate time complexity for different algorithmic parts of model, starting with the "heaviest" one: model part where saliency map is calculated.

To further decrease the absolute values of time consumption, according to [Ramík 12], we can apply a resizing procedure: for each initial input image I with size of $N = n \times m$ square pixels, pixels we define modified input image I' with size $N' = IRC^2 \times n \times m$ square pixels, where width and height depend on initial width and height through some image resize coefficient IRC.

If we use the modified image in the model instead of initial one, it can be stated that IRC value in range (0..1) can decrease time consumption of the whole model. However, the question of decrease in quality stays open; we can only operate on the level of empirical observation of dependency of several efficiency and quality metrics on the real value of this resize coefficient. Example is depicted in Figure 4.1

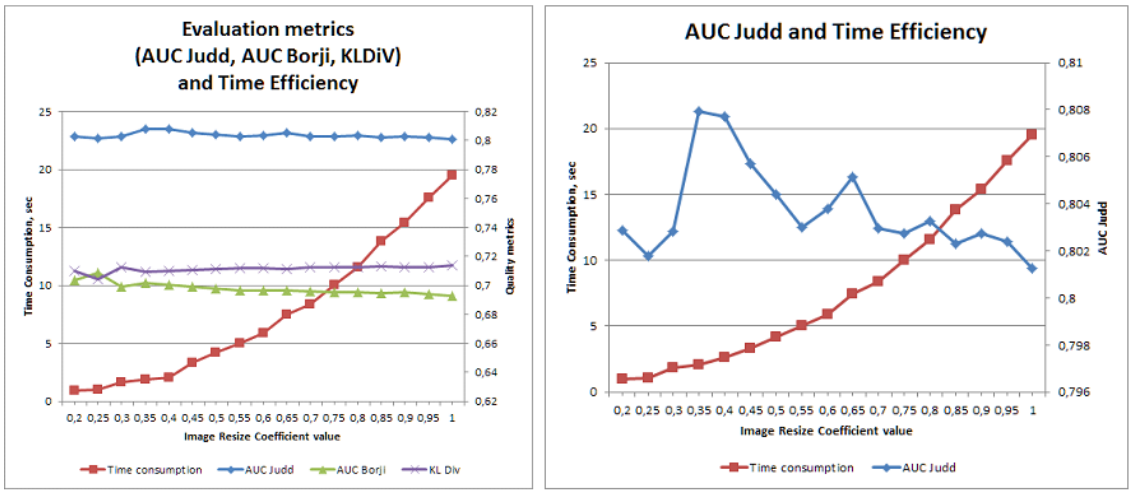

Figure 4.1: Example diagrams with average robustness metrics – time consumption and evaluation metrics AUC Judd, AUC Borji, KL Div – for different IRC values

(left diagram represents average evaluation results for model if used with a dataset with average image size 1024x768 pixels – set of 60 images from MIT1003 dataset; right diagram represents same average evaluation results of AUC Judd metric, as given in different scale). As it shows some sort of hypothetical dependency between decrease of evaluation metrics and image resize coefficient, we can only state that there could exist a range of IRC values, which gains real-time speed for some exact implementation and, depending on initial image size, can be used with just a slight decreasing effect on some evaluation metrics and even slight increasing effect on other metrics. Such "ranges of plausibleness" of IRC values can be found for each standard image size, and we provide our empirical notes of such notion in Table 4.1.

Table 4.1: Image Resize Coefficient (IRC) value ranges, depending on the robot camera frame size (resolution)

Standard Resolution Name	Size	Recommended ranges
QVGA	320x240	0.5 — 1
VGA	640x480	0.4 — 0.75
SVGA	800x600	0.2 — 0.6
XGA	1024x768	0.15 — 0.45
SXGA	1280x1024	0.1 — 0.4
UXGA	1600x1200	0.1 — 0.3
FullHD	1920x1080	0.05 — 0.25

4.3 Fire Detection Problem

The consideration of visual fire detection problem was inspired by the previous collaborations between LISSI and LSPE (Le Laboratoire Sciences Pour l'Environnement) of the University of Corsica.

A robot, equipped with VA^3V model, could provide cooperative assistance of firefighters. It is pertinent to note that the term of "cooperative assistance" is a central issue in taxonomy of the investigated system versus other fire detection approaches. In fact, the main objective in the most of the fire detection approaches (including those mentioned in the present introductory section) is to detect the presence of the fire. Often, this also means dealing with automated detection of the fire, implicitly or explicitly excluding the human operator from the processing chain.

In contrast with those systems (and the related methods), linking the notion of "firefighting assistance", the proposed approach entrench the human operator (namely firefighter) as pivotal ingredient associated to the designed system. This means that although bestowing fire detection and flame-region extraction ability, the proposed approach should also provide additional skills relating rescue of endangered individuals, at-risk within the fire disaster. In other words, such a system has to switch quite flexibly from fire's region detection to humans' detection, proffering the user (i.e. firefighter) a flexible and cooperative assistance by improving the user's awareness about the environment devastated by the disaster. Within the aforementioned point of view, we argue that proffering an artificial vision system the skill of behaving closer to the operator that uses it (i.e. proffering it a human-like conduct) is an appealing feature for raising the firefighter's efficiency in rescuing endangered people or in his (or her) firefighting action.

By acquiring some kind of artificial human-like visual attention shoving it (i.e. the system) to focus either the fire's region or the at-risk individuals, the resulted system becomes able to adapt its conduct to the firefighter's focal needs (or targets) and cooperate with him (or her) in order to achieve an improved awareness of the human operator regarding the hostile environment.

4.3.1 WiFiBot-M

Wifibot-M by Nexter Robotics[1] is suited for those who want an affordable but robust mobile platform for local surveillance. The base system is composed by a six wheel drive waterproof (IP64) polycarbonate chassis, which are controllable through

[1]Nexter Group company filiale, more info can be found at company's website: `http://www.nexter-group.fr/fr/filiales/nexter-robotics`

Wi-Fi. An example of the robot's appearance is presented in Figure 4.2[2].

Figure 4.2: Sample photo of **Wifibot-M** 6-wheels mobile robotic platform

The chassis is composed by 3 parts linked with a 2 dimensional link. It is also connectable with devices such as IP camera (MJPEG or MPEG) or any Ethernet sensor; a liteStation2 from UBNT router is the main CPU that allows data transfer, and a 5Ghz router can be added.

The instance of this robot, used for experiments in this work, is equipped with an analogue PTZ (Pan-Tilt-Zoom, three degrees of freedom) camera (WONWOO WCM-101), attached to chassis through AXIS M7001 video encoder. The robot can be controlled through Wi-Fi or Ethernet network connection.

As this particular model is definitely old (first appearance in the market was found not after 2010), it is provided with relatively slow and feeble in resources CPU; such a robot should be controlled by a remote computer, providing "external brains". In this case the WiFiBot is mostly a "camera with wheels", which is its first and main real mission – the producer claims such a robot usable in search missions or night territorial security (if equipped with a night-vision camera).

As Figure 4.3 shows, an application for robot control can be constructed. Here two parts of the system ("Camera Handler" and "Wheels Controller") represent robot-specific details of implementation – how exactly do we extract the input from the robot, and what exactly do we send to the robot in order for it to react to our algorithm.

As two other modules are more in the context of this research, the "Visual

[2]As adopted from robot's datasheet, found at `http://www.wifibot.com/download/WifibotM_datasheetEN2010.pdf`

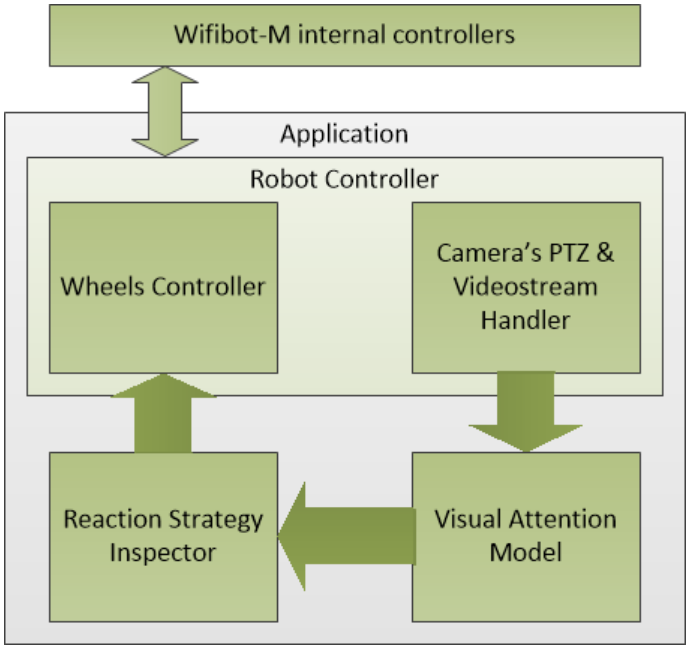

Figure 4.3: An application system around **WiFiBot-M**

Attention Model" here represents a model, needed for each task – not always we need both the bottom-up and top-down parts, as not always we need all the modules of the top-down part.

"Reaction Strategy Inspector" represents a generic approach to the input reaction. As this robot presents itself more like an agent, usable in bigger multi-agent system of real-world search, it is appealing to use here some agent-system terminology, thus making this implementation feasible for further scaling in any hypothetical multi-agent system.

As according to classical work of [Benson 93], one of the main general abilities set for a mobile robot is teleo-reactive behaviour; thus, a teleo-reactive (T-R) program is an agent-controlling program that drives him to a goal, taking into account continuously changing environment.

Such program could be interpreted as a set of productions $(K_i \rightarrow A_i)$, where K_i is an i-th condition (taken from perceptual input and stored picture of previous world states), and A_i is i-th action (which can be either an action on the world or to the model itself, or a T-R program itself). More info on formalized T-R programs can be found in [Nilsson 94], as the whole idea of T-R algorithms was coined and vastly explored by Nilsson in his works.

4.3.2 Datasets and Tuning

The validation of the investigated model in the context of fire detection has been performed on the basis of four datasets that we denote by "FireDataSet-1" (FDS-1), "FireDataSet-2" (FDS-2), "FireDataSet-3" (FDS-3) and "FireDataSet-4" (FDS-4), respectively. Three of them, namely FDS-1, FDS-2 and FDS-3, contain appraised images (evaluated by experts) and the last one (namely FDS-4) consists of not-apprised images provided by Wifibot-M; Figure 4.4 shows the robot within the above-mentioned experimental setup and gives a sample of images from FDS-4 obtained from Wifibot-M robot.

(a) (b)

Figure 4.4: Robot within the experimental setup (a) and an example of provided images (b)

First three datasets contain data both provided by LSPE and taken from different internet image storages. FDS-1 includes a subtle selection of 122 images: 61 images representing flames or wild-land fires in diverse environments and 61 images (i.e. the subset of 61 other images of the same dataset that we denote by "ground-true attention-attractive areas") corresponding to the apprised shape of flames visible in images of the first subset. The upper left-side sector of Figure 4.5 shows 12 samples of such arrangement. The upper right-side sector of the same figure gives the "ground-true attention-attractive areas" of these 12 samples. The selection criteria have been based on representativeness of the selected images regarding diversity of landscapes, colors' assortment and number of attractive objects (e.g. flames).

These examples are arranged representing an increasing processing-difficulty (four levels: "Low" labelling the lowest level and "Very-High" the most difficult one) and representing an increasing image's complexity (three levels: "One" attractive object, "Two" attractive objects and "More than two attractive objects", or just

"More").

By "processing-difficulty", we mean the difficulty of detection of the salient area (or flame) which is supposed to have mostly attracted human's attention (through his eye-fixation spots). For example, in the image representing the "unique flame visible within the quite empty land" (i.e. the first sample in upper left-side sector of Figure 4.5), the flame is clearly visible and remains the unique attractive item: this sample is labelled as "simple case" with a low ambiguity concerning the attractive item. While concerning the image representing "an aerial-view of several flames visible within the quite complex environment including a house and troubled by smoke" (i.e. the last sample), the flames are not the only salient items and part of them are buried by the smoke: this sample is labelled as "difficult case" with a very-high ambiguity concerning the available attractive item. By "image's complexity", we mean the density of salient visual information of the image. In fact, an image containing only one salient object is considered simpler regarding the above-mentioned criterion that an image including several potentially salient objects which may attract (or not attract) the human's visual attention.

In the same way, FDS-2 includes a fine selection of 110 images: 55 images representing humans in diverse wild-land fires and 55 images (i.e. the subset of 55 ground-true attention-attractive areas) corresponding to the apprised shape of humans visible in images of the first subset. Figure 4.6 shows 12 samples (upper left-side sector of the figure) and the corresponding apprised ground-true attention-attractive areas (upper right-side sector of this figure) of FDS-2, respectively. FDS-3, including 100 images following the same arrangement policy, has been built by combining part of FDS-2 with 13 additionally evaluated images provided by Wifibot-M robot.

Finally, FDS-4 includes 75 not-evaluated images extracted from video-stream sequences provided by Wifibot-M. This last dataset doesn't include any image corresponding to ground-true attention-attractive areas. Concerning images provided by Wifibot-M-based implementation, in order to make the validation scenario compatible with plausible fire-fighting circumstances, the experimental setup has been realized within the robot's scale. According to the Wifibot-M robot's size, this means some $300m^2$ area (typically $25 \times 12m^2$ action-area) and a 80-to-100 centimeters-height and 100 centimeters-width fire with smoke somewhere in that area. This also ensures a correct WiFi connection (as well regarding network connection's quality as regarding the relative simplicity of required supply deployment in outdoor conditions) and a correct energetic autonomy of the robot allowing performing several experimental tests during several hours if required. Several tests with robot moving toward and around the combustion (fire) zone with aim of detection of the fire's

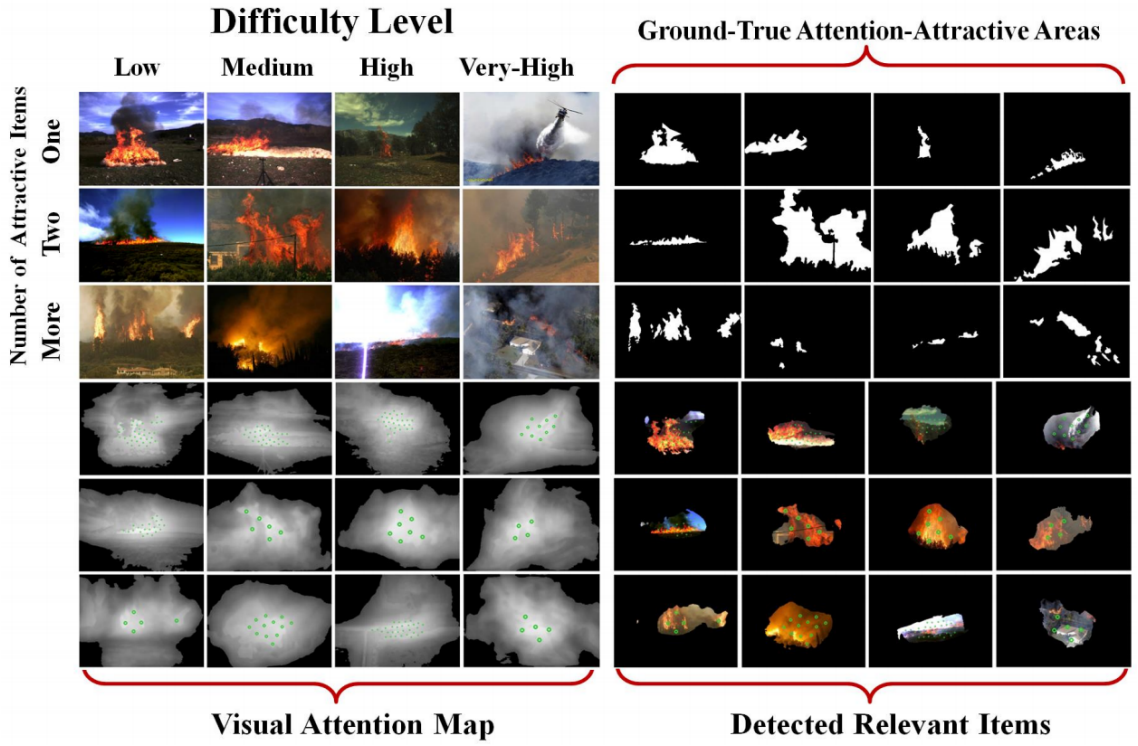

Figure 4.5: Examples of obtained simulated experimental results on FDS-1 showing input patterns (upper left-side sector), the corresponding "ground-true attention-attractive areas" (upper right-side sector), computed visual attention maps for each input image including the so-called "simulated eye-fixation meeting areas" represented as green spots (lower left-side sector) and detected relevant items in each image (lower right-side sector)

shape as salient target have been realized.

The GA-based tuning was done using FDS-1 (in order to promote fire saliency) and FDS-2 (in order to promote human figure saliency). For each tuning 3 best sets of parameters, resulting from tuning processes, have been retained. The example comparison of metrics evolution of tuned model against untuned model is depicted in Figure 4.7.

It is pertinent to notice that the AUC_{Judd} metrics dispersion characterizing the untuned saliency detection system is substantially reduced through the GA-based tuning process in the model resulting in an obvious increasing of the overall mean value of AUC_{Judd} in proposed system and thus, boosting the likeness between the model's behaviour and the human-like eye-fixation mechanism. This clearly means that incorporation of the GA-based tuning process proffers model kind of human-like artificial visual attention.

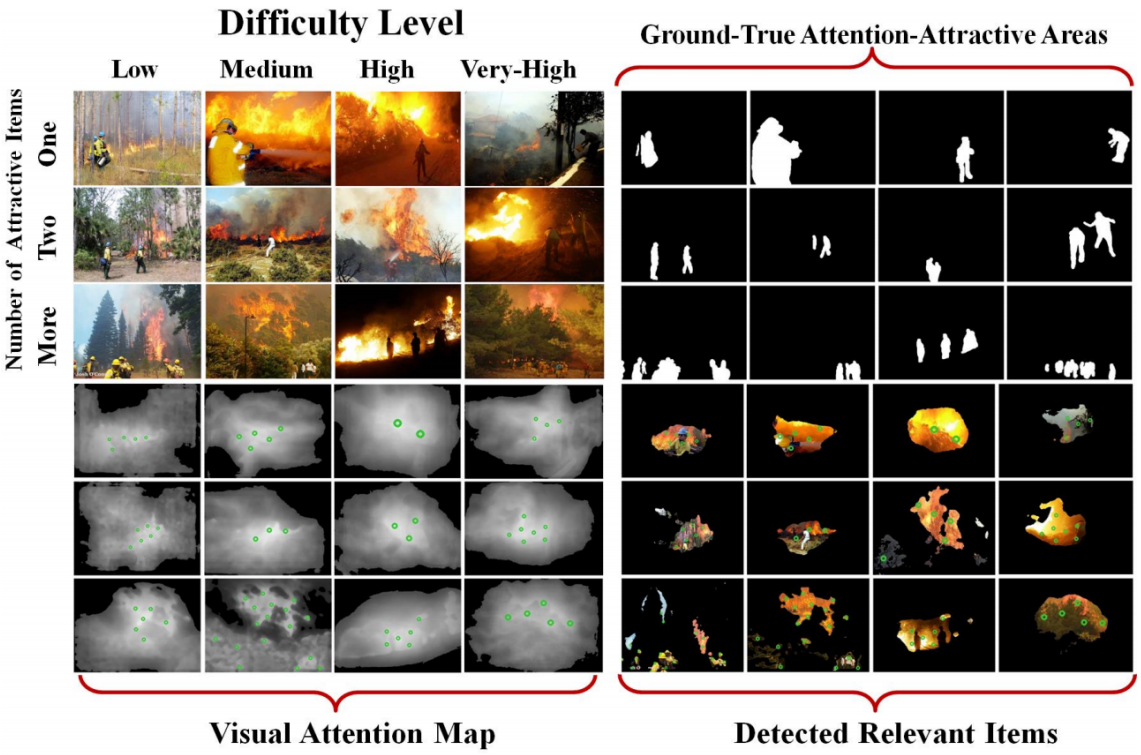

Figure 4.6: Examples of obtained simulated experimental results on FDS-2 showing input patterns (upper left-side sector), the corresponding "ground-true attention-attractive areas" (upper right-side sector), computed visual attention maps for each input image including the so-called "simulated eye-fixation meeting areas" represented as green spots (lower left-side sector) and detected relevant items in each image (lower right-side sector)

4.3.3 Experimental Runs

Using the previously-described four datasets, several evaluations have been performed involving as well tuned as untuned model:

- Evaluation-1: VA^3V model has been tuned using FDS-1 and then tested by using FDS-2 and FDS-3 ("fire tuned" model).
- Evaluation-2: model has been tuned using FDS-2 and then tested by using FDS-1 and FDS-3 ("humans tuned" model).
- Evaluation-3: All three pre-evaluated datasets (i.e. FDS-1, FDS-2, FDA-3) have been used to test performance of the raw (i.e. untuned) model.
- Evaluation-4: FDS-4 has been used as testing dataset for assessing the model's performance in outdoor conditions. This model has been tuned using FDS-1 and FDS-2.

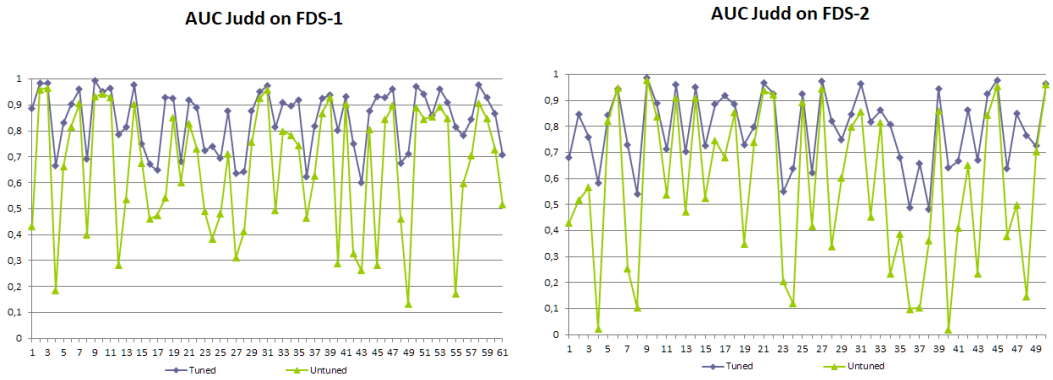

Figure 4.7: Comparison of AUC_{Judd} metrics evolution of tuned model versus untuned using FDS-1 (left) and FDS-2 (right)

We need to emphasize that first three evaluations are done as simulations, without the real usage of mobile robotic platform; while fourth is done in "real world" – when robot WiFiBot-M takes the input data from camera, and reacts according to its T-R program.

Table 4.2: Summary of obtained results for "Evaluation-1"

Dataset	Chromosome[1]	AUC_J	AUC_B	KL
	Tuning			
FDS-1	(0.26; 0.36; 99; 2.21; 82)	0.8276	0.7567	0.9926
	Testing			
FDS-2		0.8077	0.7142	1.0619
FDS-3		0.8592	0.8075	1.3628

[1] Each chromosome is represented as a set of parameters in the following order: (WSC; w_G; A; Nar; FT).

Concerning the "Evaluation-4", it is pertinent to notice that FDS-4 includes not-preevaluated data provided by Wifibot-M robot and thus, it is not possible to compute the values of the three considered indicators (namely AUC_{Judd} , AUC_{Borji} and KL_{Div}) for this set of data. That is why, the results corresponding to this last evaluation are analyzed intuitively through a visual analysis of obtained visual-attention-maps and the corresponding simulated eye-fixation meeting points' location in perceived scenery. Thus, Table 4.2 summarizes results as well for training as for testing phases relative to "Evaluation-1": reporting the winning chromosome and corresponding evaluation metrics respectively; same goes for Table 4.3 on "Evaluation-2". Finally, Table 4.4 recaps results obtained for raw system, where the values of parameters were randomly initialized and not tuned.

These evaluations show, that the model acquires a kind of "human-like" staring

Table 4.3: Summary of obtained results for "Evaluation-2"

Dataset	Chromosome	AUC_J	AUC_B	KL
	Tuning			
FDS-2	$(0.14, 0.1;\ 0.28; 23;\ 1.5;\ 37)$	0.7895	0.6863	1.3115
	Testing			
FDS-1		0.7291	0.6940	1.4478
FDS-3		0.8494	0.7389	1.3978

Table 4.4: Summary of obtained results for "Evaluation-3"

Dataset	Parameters	AUC_J	AUC_B	KL
FDS-1	$(0.25, 0.29;\ 0.25; 17;\ 0.68;\ 47)$	0.6397	0.7207	1.3013
FDS-2		0.6519	0.6656	1.8773
FDS-3		0.7802	0.7899	1.3686

skill of the inspected landscape within the context of confused and hostile environment of the fire disaster. This is achieved through saliency detection and artificial visual-attention proffering such a system the ability of extracting autonomously flame regions from raw camera images, the capability of detecting autonomously individuals (humans) at risk within the flames regions from those raw camera images and the skill of cooperating (i.e. jointly inspecting) with the firefighters (i.e. users) in order to improve their awareness about the concerned disaster.

Other existing systems are concerned either by fire's shape detection or by humans' tracking, but don't deal with both of them within hostile and confused wildland fires context. Moreover, as far as we may know, there is not a system which takes aim cooperating with the human-operator: whole other systems are "just tools used by the operator".

For this reason a comparative study between the proposed tuned model and other existing approaches will only show that some "specialized algorithms" will perform as good or even better in the task for which the concerned algorithm has been designed. In other words, many dedicated algorithms execute faster and better, than the human being, the specific task for which they have been designed, but they are not able to deal with or adapt themselves to various situations for which they haven't been designed. Consequently, an accurate comparative study remains difficult to realize, because still there is no equivalent systems or comparative implementations.

However, in order to give an overall idea of the investigated model's performance in this context, Table 4.5 provides an F-measure-based comparison between the proposed system and the BMS algorithm (proposed in [Zhang 13], already mentioned in

section 1.3). This algorithm is used as well for complex object's recognition as an eye-fixation predictor, and thus somehow fits similar tasks. As the table shows through the considered metrics, although reaching somehow comparable performances for FDS-1 (while better), the investigated system outruns BMS in detection task when the inspected landscape includes both humans and flames.

Table 4.5: Summary of obtained results for F-measure-based comparison between BMS algorithm and the VA^3V model on three datasets

F-measure	FDS-1	FDS-2	FDS-3
BMS	0.3414	0.1186	0.1578
VA^3V	0.4456	0.2082	0.4097

Figure 4.5 and Figure 4.6 provide example of obtained experimental results relative to "Evaluation-1" and "Evaluation-2", respectively. The upper left-side sector in these figures shows 12 samples of images of each dataset, representing increasing processing complexity (relating as well the number of salient items as the mistiness of the seeming environment). The upper right-side sector in these figures shows the "ground-true attention-attractive areas" corresponding to the illustrated 12 samples. The lower left-side sector of each figure shows the computed visual attention maps for each illustrated image including the so-called representative points (shown as green marks), introduced and defined in subsection 3.3.6. Finally, the lower right-side sector of each figure provides the relevant items detected by VA^3V model and the simulated eye-fixation meeting spots matching up the detected items.

Figure 4.8 provides examples of experimental results relative to "Evaluation-4", obtained using Wifibot-M robot in outdoor environment. The left-side results have been obtained from tuned VA^3V model and those visible at left-side are resulted from the rough system set by arbitrary parameters.

Several observations may be formulated regarding the reported results. The first remark relates quantitative results summarized in Table 4.2, Table 4.3 and Table 4.4. Those results show that the tuned system admits discernibly higher values for AUC_{Judd} indicator: around 0.8 (or higher) for tuned system and around 0.6 for system operating with arbitrary parameters. Let us remind that the highest attainable score for AUC_{Judd} indicator is 1 while, a uniformly random saliency map conducts to the score of 0.5.

This means that such a measure evaluates the assessed technique's quality versus the random process: scores reaching values below 0.5 are interpreted as "worse than random process". Thus, based on the scores reached by arbitrary set system and those reached by tuned system, it becomes perceptible that the arbitrary set system

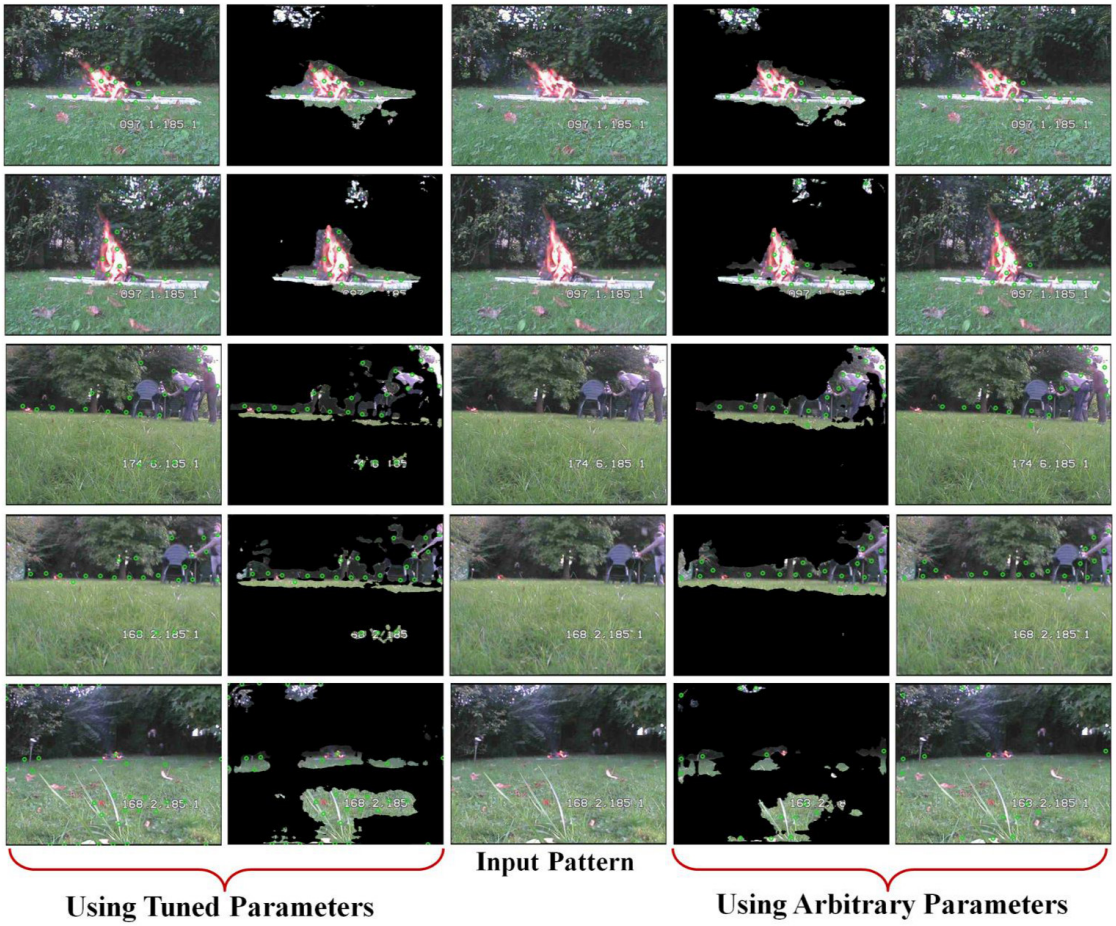

Figure 4.8: Examples of experimental results relative to "Evaluation-4", showing detected items achieved by tuned system (left side) and those obtained when the model's parameters are arbitrary set (right side)

operates closer to a random process than the tuned one and thus, its scrutiny of the landscape appears as more unsystematic than a visual-attention-based search of same items.

The other comment relates experimental results highlighted in Figure 4.5 and Figure 4.6. In fact, one can note that, even though more dispersed in images resulting from untuned system, simulated visual-attention meeting spots match quite fittingly the detected items when the input image is unchallenging. If the untuned system carries out somehow comparable saliency detection for low-complexity landscapes, however, it often fails in detecting accurate items (i.e. either flame or humans) when the input image become complex or ambiguous. This may be explained by the fact that although operating with inappropriate parameters, the basic saliency detection mechanism remains enough for extracting salient items within smooth conditions. However, within degraded or tricky conditions, additional visual attention is needed

for detecting the expected items.

Finally the last remark relates experimental results illustrated in Figure 4.8. Here also, the tuned system confirms its advantage versus the un-tuned one, especially when either the landscape is more complex or when the flame is confusing. Results show various examples, from simpler cases (firsts rows of the figure) to critical ones (lasts rows of the figure). The two firsts correspond to a quite comfortable situation where the flame is the most salient event within the inspected landscape. In fact, although the model's attention is clearly and effectively focused on flame, here the saliency detection of the raw untuned model detects quite correctly the unique salient flames. Contrary to those simpler cases, other examples illustrate more critical situations where the landscape contains either both humans and flames (as in two next rows) or the flame, though unique salient event, remains far from the robot making other closer objects (as flowers, etc.) as salient as the flame. In those cases, as the figure shows, the detection of target objects (i.e. flames and humans) is visible for the tuned system but not achieved for the unrefined system. Alongside validating, the investigated concept and its implementation on Wifibot-M robot within realistic outdoor conditions, the obtained results confirm the generalization ability of VA^3V model. These results show also the efficiency and the pertinence of the incorporated GA-based tuning strategy.

4.4 Arbitrary Attention Problem

While WiFiBot-M provides pretty interesting possibilities in terms of mobility, the bigger context of the work was to provide the means for improvement of general social skills of any autonomous mobile robot. Due to this another problem context was formulated – a humanoid robot should be able to solve visual search problem, depending on the combined visual attention model.

4.4.1 Humanoid Aldebaran Robots

To be able to apply the problem to humanoid robotics, we need to provide several experiments exactly on these humanoid robots. This has been done on the robots, created by SoftBank Robotics[3].

Humanoid robots NAO and Pepper were designed to make the interaction with human beings as natural and intuitive as possible, as equipped with multimodal

[3]formerly known as Aldebaran Robotics, but acquired in 2015 by SoftBank. More info can be found at company's website: `https://www.ald.softbankrobotics.com/en`

sensoring as well as some basic autonomous life modules. (visuals shown in Figure 4.9[4]).

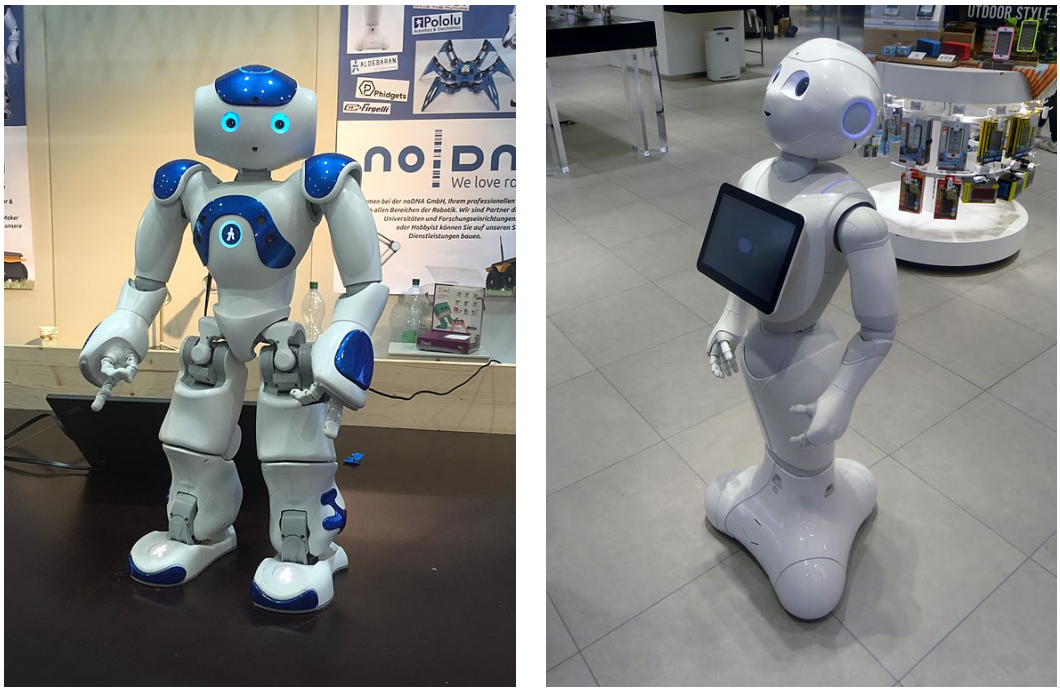

Figure 4.9: Physical overview of the robots **Aldebaran NAO** (left) and **Aldebaran Pepper** (right)

The NAO robot is "older brother" of Pepper, and is equipped with less productive hardware: Intel ATOM Z530 1.6GHz, camera MT9M114, VGA@30fps, 72.6°DFOV (60.9°HFOV, 47.6°VFOV). With also having the height of the robot only 58 cm, it is rather difficult to scale the objects in a controlled experimental environment.

The Pepper robot, on the other hand, being more recent development of the SoftBank engineers, is much more "humanoid" and is equipped with better hardware: Intel Atom E3845 Quadcore 1.91 GHz, 2 cameras OV5640 which can supply VGA@30fps or up to 4VGA@1fps, 68.2°DFOV (57.2°HFOV,44.3°VFOV). Also, the robot is equipped with ASUS Xtion 3D sensor, able to provide data in focus range 40cm–8m with 320*240@20fps, 70.0°DFOV (58.0°HFOV,45.0°VFOV). While the multi-sensor analysis is not in the scope of this work, we have to consider auxiliary usage of the 3D sensor in order to be able to provide a navigation-like control of the robot.

Another advantage of the Pepper robot is its dimensions: it has 1.21 meters height, which gives us more correct and human-like scales for the experimental setups.

[4]Images adopted from respective Wikipedia pages, `https://en.wikipedia.org/wiki/Nao_(robot)` and `https://en.wikipedia.org/wiki/Pepper_(robot)`

Both robots work on the Linux-based NAOqi operational system, which provides also NAOqi API to all the firmware modules (both considering hardware and pre-installed software considering the previously mentioned "autonomous life modules").

As Figure 4.10 shows, to be able to use the Aldebaran humanoid robots, we need to create both the intermediate application tier which contacts with robot through NAOqi API, and the "client tier" which implements the combined visual attention model.

In this scheme we use the input from robot's camera (or cameras) as standalone frames in order to use the high-level resolutions. Then these frames, processed by *ALVideo* NAOqi module and through intermediate tier of NAOqi API, are given to be processed by both bottom-up and top-down modules of the combined model.

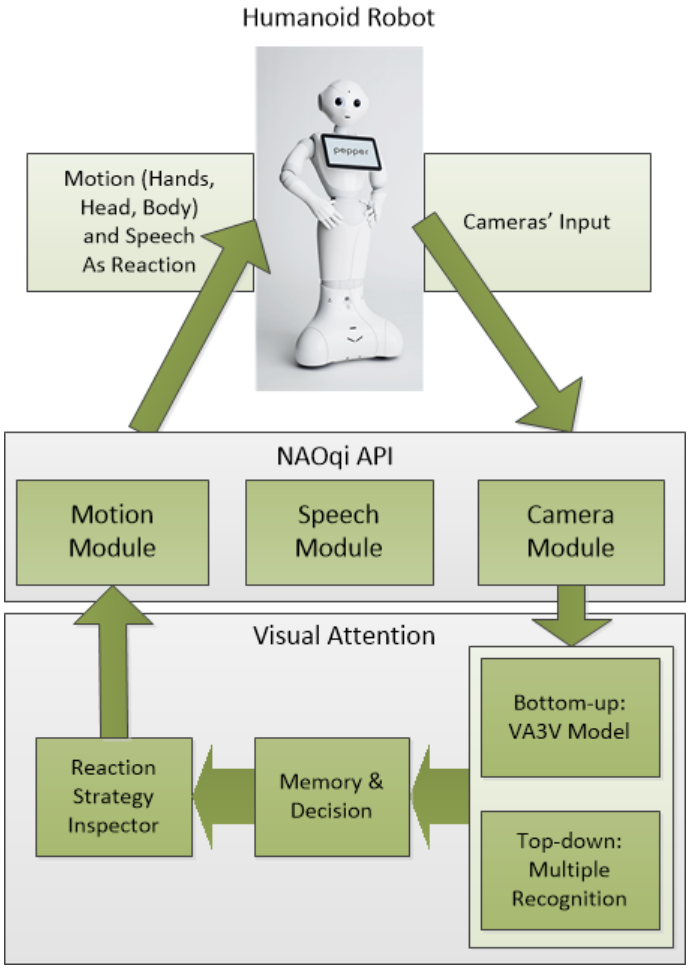

Figure 4.10: An application system around NAOqi-based humanoid robots

Depending on the task, given to the robot, the decision module provides the attention model results to the reaction strategy inspector, which decides on the next robotic reaction. Thus, if the task implies, that robot should point by its hand onto the most representative points, the strategy inspector should provide this

information.

The reaction inspector creates the reactions, which are then fulfilled by robot's motion and speech modules (or the robot does nothing, if there are no valid reactions at this iteration).

4.4.2 Hand Pointing Problem

Let us assume a visual attention problem concentrated on full attention mechanism and robotic modules' usage cycle: *The robot watches in front of itself, and the most salient representative point of each iteration is found; then robot reacts with pointing at this point by hand and saying the location of this point.*

To be able to assess the results of such an experiments, we have to either use an empty controlled environment where ground truth is easily generated, or use existing complex images from eye fixation benchmarks. Assuming, that the "brightest" pixel (in terms of grayscale ground truth maps, meaning the highest probability of eye fixation at this point) is the most "eye fixating", we use the second approach in this experiment.

4.4.2.1 NAO with VA^3V Model Experiment

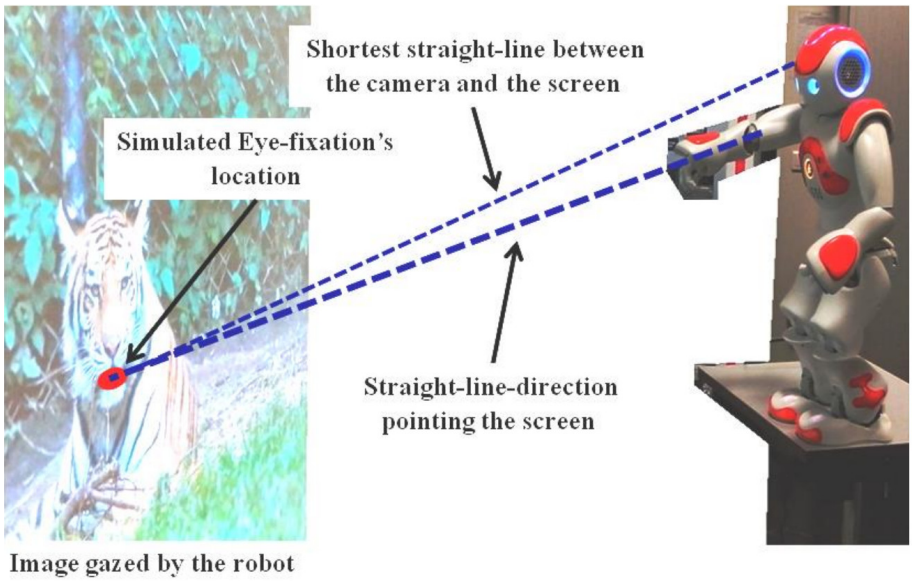

Figure 4.11: Experimental setup with NAO watching an image projected on the screen

Firstly the experiment viability was tested on the NAO robot with only the bottom-up module enabled (thus, the top-down and decision parts are not being tested at this moment). The robot has been placed in front of a screen and a

subset of images, from the testing and validation datasets T-60 and V-60 (mentioned and detailed in subsection 3.3.5), have been projected on the screen allowing the robot watch them one-by-one; Figure 4.11 illustrates the experimental setup (NAO watching an image projected on the screen). While the NAO's camera isn't a moving camera, remaining far from the sophistication inherent to the human's eye, the robot's eye-fixation could be simulated by the straight-line-direction pointed by NAO's arm in such way that the corresponding impact-point (location on the screen symbolized by the red area) corresponds to the shortest straight-line between the NAO's camera and the screen when robot's face is parallel to the screen.

The reported samples (illustrated in Figure 4.12) correspond to two results obtained from two experimentations: the first group (e.g. the two left-side columns) relates the robot's VA^3V-based visual behavior with tuned parameters, while the second (e.g. the two right-side columns) corresponds to robot's visual behavior with model's parameters set arbitrarily. Both two figures illustrate images provided by the robot's vision system (i.e. what NAO sees) where simulated eye-fixation points (also called representative points as given in subsection 3.3.6), conformal to the above-mentioned definition, are reported as "green marks". The ground true eye fixation areas corresponding to humans' visual mechanism are reported in the middle column.

It is pertinent to note the higher accuracy of simulated eye-fixation points' distribution regarding the ground true eye fixation area (matching the ground true eye fixation areas in both levels of difficulty), when the robot's VA^3V-based vision operates using tuned parameters. In contrast to this, robot's VA^3V-based vision operating with arbitrary-set-parameters fails the salient object's detection, except for firsts images (e.g. the three firsts rows) where the complexity is low.

It is also pertinent to note the emergence of some kind of robot's human-like visual behavior making its perception of the surrounding environment closer to the human way of perceiving that same environment. In fact, an indicator stressing such emergence could be defined on the basis of true positive and false positive rates, denoted TPR and FPR, respectively.

Here true positive rate represents the rate of green marks, which are **correctly** in the region of ground truth, against the whole number of the green marks produced for this image. False positive rate uses the same logic, being the rate of **misplaced** green marks (not in the region of ground truth) against the whole number of the green marks.

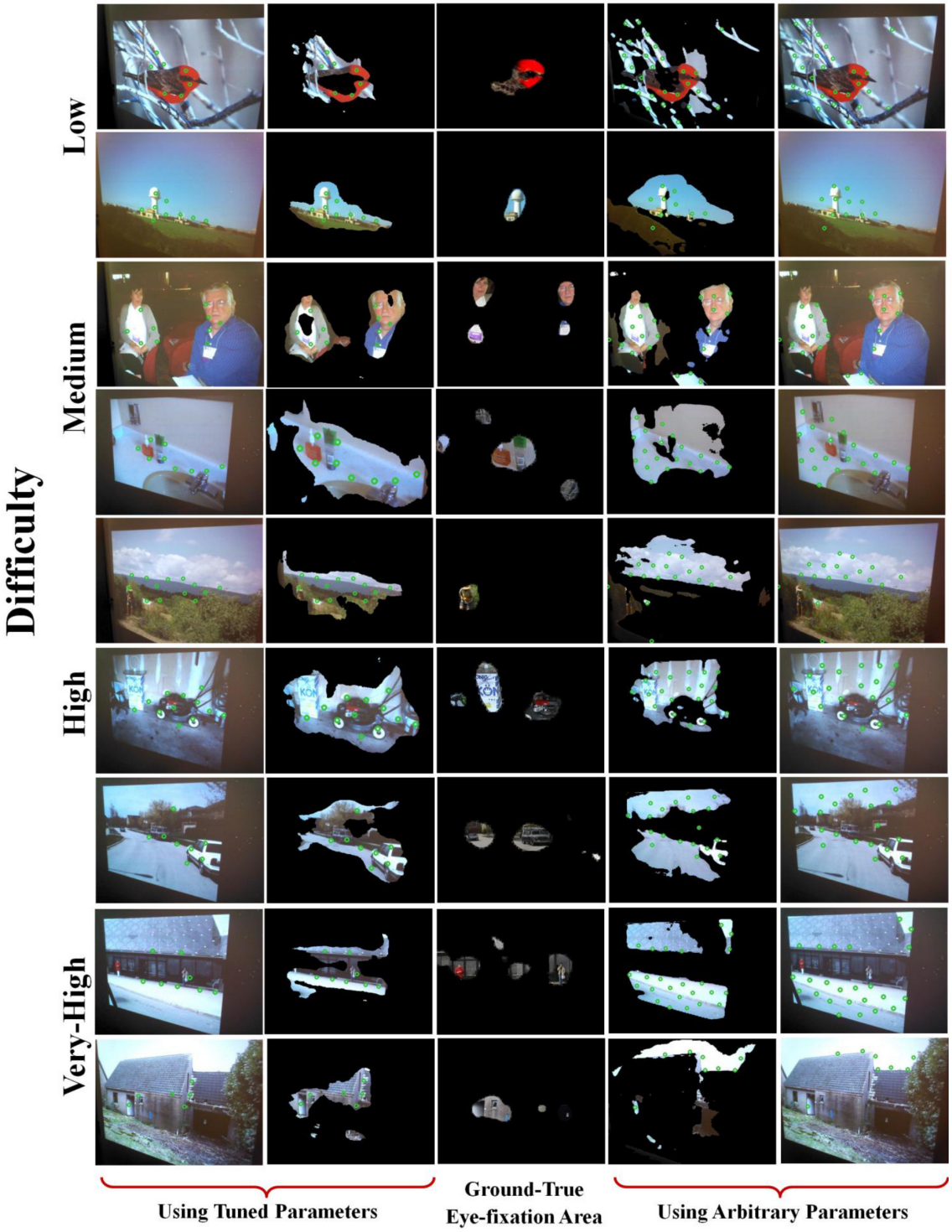

Figure 4.12: Examples of experimental results for patterns from V-60 dataset showing robot's eye-fixation behavior when VA^3V uses tuned parameters (left-side results) and when the model's parameters are arbitrary set (right-side results)

Based on these rates, one can emerge an indicator "Concentration rate" (high-

lighting some kind of machine's attentiveness degree), and we define it as shown in Equation 4.1.

$$\chi = \frac{TPR}{FPR} \tag{4.1}$$

Taking into account the aforementioned definition, one can define $\chi_{k\%}$ as the smallest value of χ corresponding to the true positive patterns where at least k percent ($k\%$) of their simulated eye fixation points (i.e. "green marks") match with the corresponding ground-true eye-fixation-maps. Let $TP_{k\%}$ denote the number of true positive patterns with $\chi \geq \chi_{k\%}$. E.g., $\chi_{30\%}$ corresponds to the smallest value of the above-mentioned indicator relating the true positive patterns including at least 30% simulated eye-fixation points matching with the ground true eye fixation map of the concerned patterns.

A comparison of $TP_{k\%}$ values, corresponding to the VA^3V model, operating with tuned parameters, with arbitrary (e.g. not tuned) parameters, and a random (Monte-Carlo style) process is shown in Table 4.6 for $k = 50(\chi_{50\%} = 1.0)$ and $k = 30(\chi_{30\%} = 0.5)$.

Table 4.6: Comparison of $TP_{30\%}$ and $TP_{50\%}$ for VA^3V model operating with tuned parameters, VA^3V model operating with arbitrary (e.g. not tuned) parameters, and a random Monte-Carlo style process

	VA^3V tuned	VA^3V untuned	Random
$TP_{50\%}(\chi \geq 1.0)$	23%	3%	0%
$TP_{30\%}(\chi \geq 0.5)$	58%	6%	0%

In fact, as it is visible from this table, the number of patterns for which the value of exceeds 0.5 or 1.0 is significantly higher for VA^3V model operating with tuned parameters. In other words, the number of patterns matching the human's eye-fixation mechanism increases significantly for VA^3V model operating with tuned parameters making the robot acquiring a kind of human-like visual behavior.

4.4.2.2 Pepper with Full Model Experiment

In terms of full model this task is categorized as "free visual search", meaning that importance is given only to specific pattern-based objects, like human faces and writings (as shown in Equation 4.2). Also, we have to "disable" the episodic memory analysis of the visual context, as the robot doesn't move.

$$I(''HumanFace'') = 1,\ I(''Text'') = 1,\ I(Other) = 0,\ I_{img}(k) = 1,\ \forall k \tag{4.2}$$

In this modification of the initial experiment, we place the robot exactly in front of the display, so that the visual field of the robot almost perfectly corresponds to the display.

Firstly, the robot hand-pointing is calibrated to 8×8 grid on the display, so that the hand-pointing region corresponds to one of the cells of the calibration grid each time, that robot tries to hand-point anywhere (scheme shown in Figure 4.13). In this case anyone from aside can quickly understand whether each iteration provides good or bad results.

After the calibration, the subset of images is shown at this display (the same set, as in previous experiment, T-60 and V-60). The robot watches them, again, one-by-one.

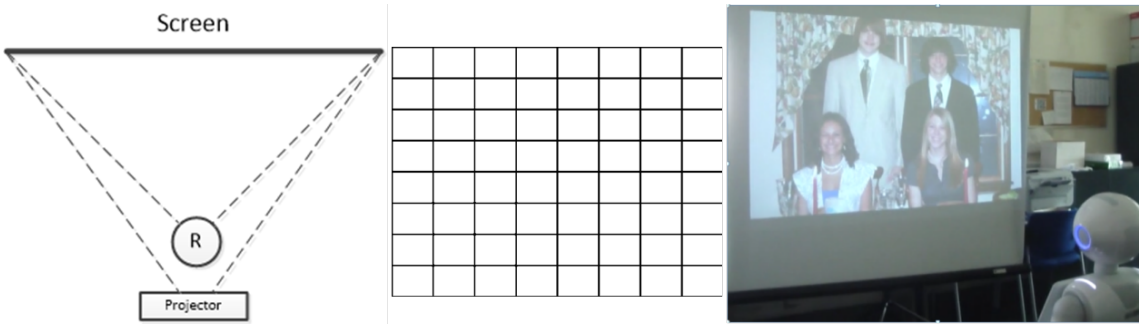

Figure 4.13: Pepper robot hand pointing experiment, from left to right: overall scheme of the experimental environment, view from above; calibration grid, used for more precise hand pointing; example photo depicting the robot (lower right corner) watching the display

Due to the setup modification, we are able to apply different (from previous setup) evaluation means to this experiment. E.g., as the visual field of the robot and display almost perfectly match, this means that the image, produced by robot, almost matches geometrically to the initial input image. This leads to usability of all the eye fixation metrics, based on the comparison of initial ground truth grayscale maps against the visual attention maps, produced by the attention model.

Thus Figure 4.14 depicts, as an example, the robot's vision, the saliency maps, produced by BU unit and by combined model, as well as the example of text recognition.

The evaluation metrics are shown in Table 4.7, depicting the amelioration of the evaluation metrics for combined model against the only bottom-up approach.

As it is visible from Table 4.7, the AUC_{Judd} and AUC_{Borji} metrics become slightly better for the combined models against the sole bottom-up models. While this enhancement doesn't look drastic, it is pertinent to remember that the top-down

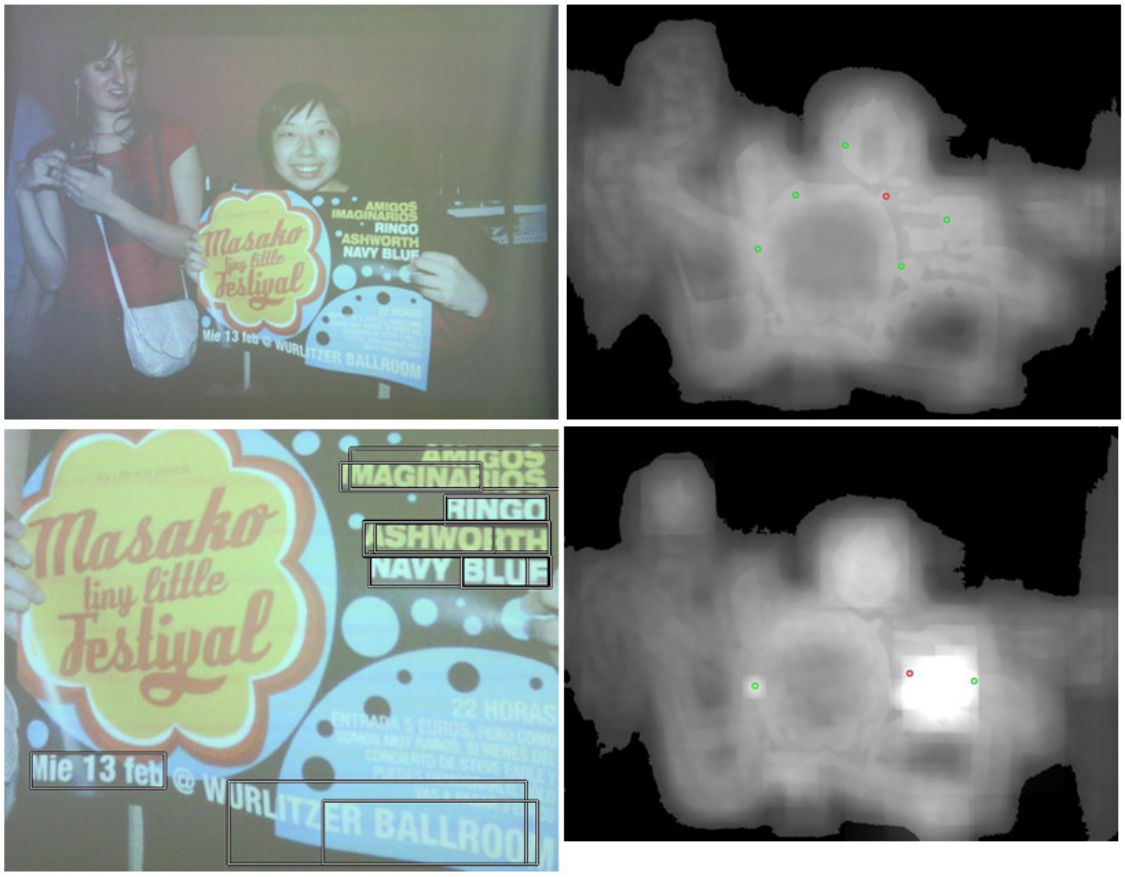

Figure 4.14: Pepper robot hand pointing experiment, from left top corner clockwise: displayed image as perceived by robot; saliency map as given by only BU unit ($M_{VAM}(k)$); saliency map as combined result of both BU and TD units ($M(k)$); crop of the input image with gray rectangles depicting correct text recognitions

unit has a little different destination, and saliency maps improvement is, in some way, a secondary target, which is yet achieved.

Also we should point out the deterioration of Küllback-Leibler divergence due to the top-down unit generating additional information, thus increasing entropy and the divergence.

4.4.3 Arbitrary Item Search Problem

Let us assume a more complex visual attention problem, which also somewhat falls into the firefighting context of the first experiment: *The robot has to find fire extinguisher, which is situated somewhere in the building.* As this experiment adds the necessity of navigation-like behaviour, we enable all the modules of the combined visual attention model, as well as some additional modules concerning pseudo 3D-vision in order to be able to find open space and calculate the needed distance to

Table 4.7: Comparison of VA^3V model against combined model operating with parameters, tuned over T-60 dataset, on evaluation metrics over two different datasets

Model	Dataset	AUC_{Judd}	AUC_{Borji}	KL_{div}
VA^3V_{T-60}	V-60	0.832	0.645	0.613
$VA^3V_{T-60} + TD$	V-60	0.856	0.672	0.901
VA^3V_{T-60}	V-523	0.831	0.652	0.59
$VA^3V_{T-60} + TD$	V-523	0.852	0.675	0.983

move.

It is pertinent to note that the robot navigation problem itself is a huge research field, which is totally out of the scope of this work. The "navigation"-like module, which we use here, is more a set of reaction rules, where robot moves, yet neither it creates a virtual map of surroundings, nor calculates optimal paths.

Experimental setup is depicted in Figure 4.15; the robot is shown as a triangle with letter "R", with the top corner showing the direction of robot's initial vision (not catching the evacuation door). Human in the room is shown as a circle with letter "H". The room, where the robot starts, is not exactly "empty": it contains several objects, both known and not known to the robot, which are not of the interest for it in this particular task. If the robot doesn't see the exact item of interest, it turns around for 60 degrees, and starts new iteration. When the robot sees the already seen scene, is starts to search for evacuation door in order to get out of the room where it has already seen everything (the importance calculation is shown in Equation 4.3).

$$SpecObj = [Text, FireExtinguisher, ExtinguisherSign, EvacuationDoor, DoorKnob] \tag{4.3}$$

$$I(Obj) = \begin{cases} 1 & \text{if } Obj \in \{SpecObj\} \\ wup(SpecObj, Obj) * CL(Obj) & \text{if } Obj \notin \{SpecObj\} \end{cases} \tag{4.4}$$

When the robot finds the evacuation door, it analyzes whether it could come through it (via 3D-sensor). If not, the robot tries to open it (with fallback scenario as asking human to open the door, if a human has been seen previously, else the algorithm ends as failed), and when the door is open, - the robot goes out of the room, where it starts the whole same scenario again.

In such an experiment the direct assessment via previously mentioned metrics

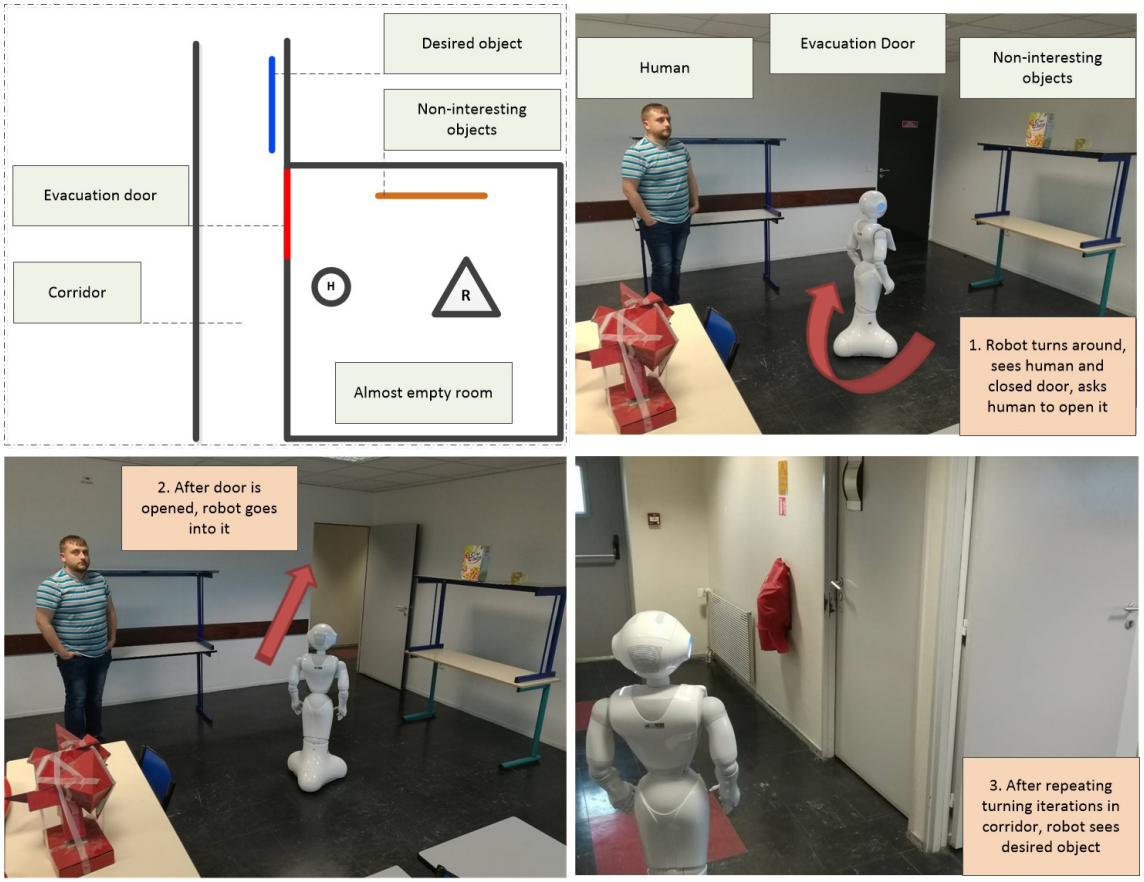

Figure 4.15: Arbitrary item search experiment scheme. Left-upper picture represents the schematic overview of the experimental setup; other pictures briefly depict the algorithm of the robot's actions based on the real photos

is not really possible due to non-existence of ground truth. While in the "Hand-pointing experiment" the ground truth existed, it was just not applicable in automatic way, so the "green point based" assessment was possible; this experiment, on the contrary, could work only as "proof-of-concept", as if the whole model can be used for such reactive behaviour of the robot, and if it is possible to further complexify the behaviour and mix the visual attention model with different other modules in order to proffer the resulting system on the next level of human-inspired social behaviour.

The most important results of the experiments are shown in Figure 4.16. First column shows the original input images from the most important iterations - robot sees non-interesting but known objects, robot sees human, robot sees evacuation door, robot sees desired object.

Second column represents the input images with the bounding rectangles, depicting the different recognition results.

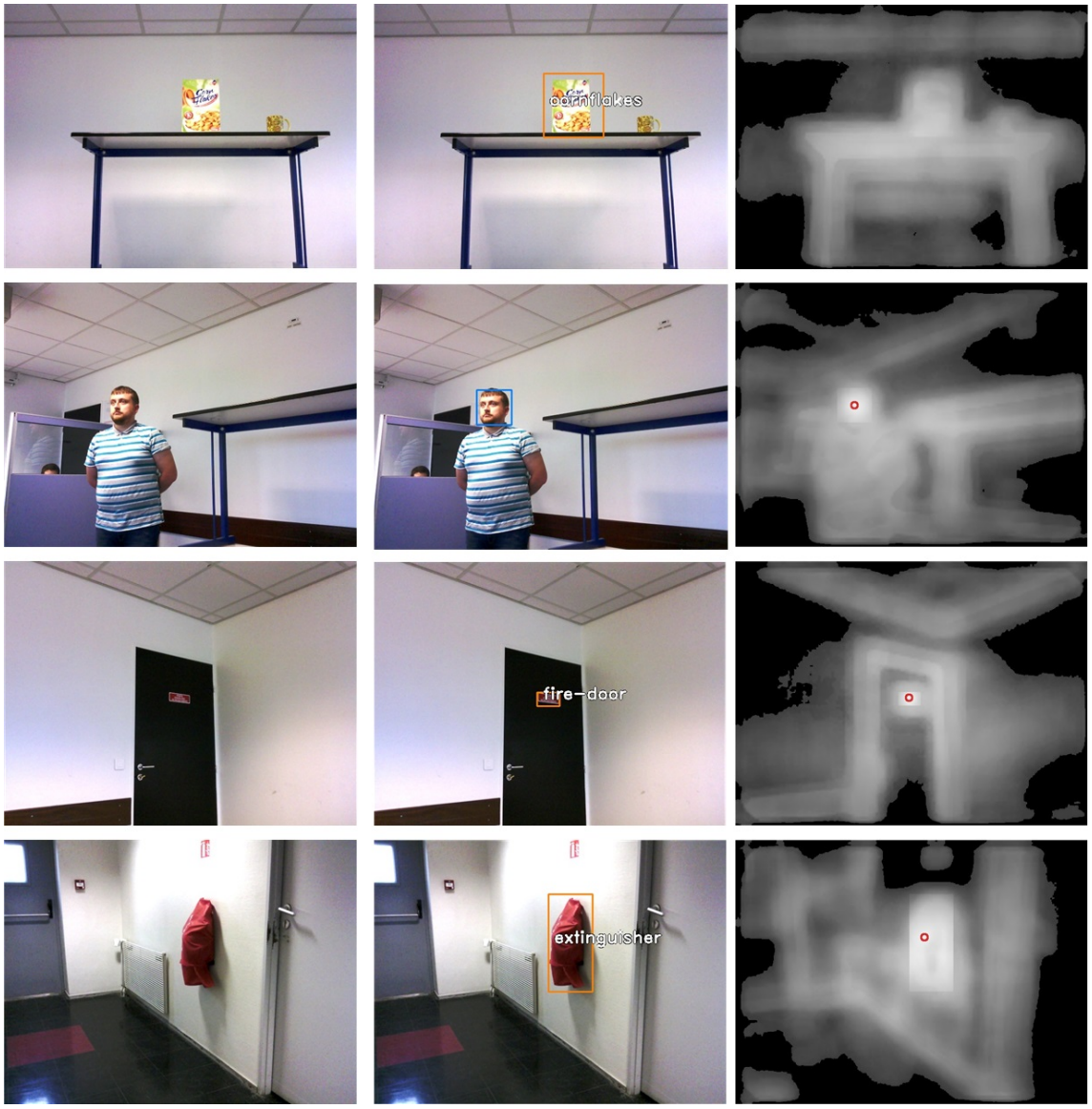

Figure 4.16: Input images from robot's camera in the experimental setup, from left to right: the original images, the images with recognized objects bounded, and the final mixed BU/TD saliency maps

Orange rectangles show recognition of known objects by keypoint-based BRISK algorithm, and blue rectangles show recognition of faces by pattern-based Viola-Jones framework. The results of letter recognition and broad recognition by CNNs are not given, as this experiment did not include any object, "true positively" classifiable by them; thus, a red chaotic object in left corner of the overview pictures (Figure 4.15) cannot be classified even by humans either than "an object of art", while tables, chairs, heaters are not really recognizable objects. Also in this experimental setup there was no texts with letters big enough for robot to be able to

"read" it.

Third column shows the final saliency map, mixed already from bottom-up saliency map and top-down importance map. Red point represents the most important object. As the first image does not hold an important (according to the task) object, there is no addition of saliency intensity in the first saliency map, as well as there is no red point of importance.

The results of the experiments give us a moderate justification to say that our combined approach to artificial visual attention can be applied in more complex problems of autonomous robotics vision, if used with other modules and approaches, such as navigation, high-level reactions, etc. It is also pertinent to notice, that while in this experiment the robot shows an ability to navigate itself, the approach is in no way a substitution for the fully dedicated navigation module; the ability for navigation in this experimental setup is provided by two additional modules:

1. teleo-reactive approach with pre-given conditions, such as *"IF there is no desired object AND a door is found AND a human is found, THEN ask human to open the door, wait, go through the door"*;

2. Pseudo-3d infra-red depth sensor ASUS Xtion 3D, used by the robot to extract the depth information, which is then processed by pseudo-3d-vision module (as firstly shown in [Fraihat 15]). This data provides a possibility to understand, whether the door is really closed, or already opened, and what is the exact distance which the robot could go for through this door.

4.5 Conclusion

The purpose of this chapter was to "close the loop" of the overall design for the combined model, presented in this work, as well as show the implementations on several robots, the simulations and the experiments, done throughout the working process, – "proof of concept", so to speak, for usage of the different modules as well as the combined model in whole.

The reader was first familiarized with several in-depth notes on the implementation details. Then, in next sections, we show the experimental setups and protocols, ordered by ascending complexity, as well as the mobile robotic platforms used in this setups, – Nexter Robotics WiFiBot-M, Aldebaran NAO, Aldebaran Pepper. Also the overall implementation schemes, used for these robots, were shown in order to explain the mode of execution for these experiments.

In this chapter all the elements of the combined model were put together in general visual attention system implementation for the Aldebaran humanoid robots.

The system in its complete state has shown that it is capable of fulfilling the tasks for which it has been designed, and complex attention-based behaviour emerged in the robot during the experiments.

The experiments, most notably the last one, underline an important detail of this system: it can be used in ensemble with other modules, such as navigation, depth vision, high-level decision making, or hearing. Such an ensemble can provide much more complex behaviour, giving the robot an ability to solve more sophisticated problems in autonomous or semi-autonomous ways.

General conclusion

Conclusion

In this thesis, we proposed and accomplished a combined visual attention model by the conditions posed in the field overview paper by Borji & Itti [Borji 13a], including bottom-up and top-down directions of attention, as well as decision module which completes the combined attention module.

Nowadays there exists only narrow context of the problems which are solved in this research domain: image treatment, such compression, thumbnailing or segmentation; object detection; advertising; several medical-context visual applications. Giving an overview of existing solutions, the bibliographical study has notably pointed out problems related to previously proposed solutions – lack of combined visual attention models, which could handle different, more general tasks. Such a study has allowed for an objective evaluation of achievements concerning autonomous visual attention in machines as well as an outline of open problems currently existing in this domain.

According to the results of this study, we propose a more-general, combined artificial visual attention approach, which is suitable to be applied in artificial vision problems on the mobile robotic platforms, either autonomous or semi-autonomous. The overall contribution and conclusive remarks on the suggested approach could be summarized as follows.

First, we develop the bottom-up model based on several bio-inspired concepts, based on a continuation of works of both the LISSI laboratory and third-party laboratories (most notable – [Liu 11] and [Ramík 12]). This model has shown itself on the comparable level with the state-of-art saliency models in terms of most evaluation metrics used in the field, being a little set back by the convolutional neural network-based methods. Yet, the classical contrast-based approach which lies in the base of our model provides less algorithmic complexity ($O(N * \log(N) + N) \leq O(M * P * I + P * I \log I + M * I * \log I)$, which implies at least 50% higher speed of processing if correctly implemented) and thus is able to create more faster results with only several percent of difference in main evaluation metrics (NSS, AUC_{Judd},

AUC_{Borji}, KL, $sAUC$).

Second, we propose the genetic algorithm-based approach to tune this model, in order to use it with different sets of parameters for different task contexts, as well as a novel approach for evaluation of the visual attention models in real world experiments via representative points, which (as a side effect) could also be used after some tuning as a quasi-saccade model, e.g. usable in artificial eyes for better results against the uncanny valley problem. The simulations, as well as quasi-real-world experiments on mobile robotic platforms (Nexter Robotics WiFiBot-M and Aldebaran NAO), show the ability of generalization as well as emergence of human-like attention in the model due to genetic tuning. Also, different sets of parameters provide different attention-like behaviour, which means a hypothetical possibility to interswap the models among different tasks to be able to achieve more complex results, such as the fire detection models used as a backup attention in object search tasks in order to provide the "background attention" for the emergency cases.

Third, we outline a structure for combined visual attention model, grouping the previously mentioned bottom-up model with several state-of-art object recognition techniques, constituting the top-down attention part of the combined model, as well as decision module represented in the structure of Baddeley-Hitch working memory schema (as firstly given in [Baddeley 74] with additions, given in [Baddeley 00]). This system stands out from similar existing algorithms as it has more complex bio-inspired structure, and can work as a basis for the further complexification by adding new components (and changing the obsolete ones), while providing comparable by efficiency results in real time, if compared to other state of art approaches. This combined model is then also validated via experimental setups using Aldebaran Pepper social humanoid robot. These experiments show additional abilities, provided for the robot, if the combined approach is used in ensemble with other third-party modules such as pseudo-3d-vision, navigation, high-level decision making, hearing.

Fourth, in order to provide the aforementioned simulations and experiments, we create several image datasets, either derived from existing benchmarks (T-60, V-60 from MIT1003 [Judd 12] and Toronto [Bruce 07]), from mix of robotic visual input and existing close-source datasets (fire and humans images in datasets FDS-1, FDS-2, FDS-3 and FDS-4), or completely created by us (Things-50, more details in Appendix A). Also we have implemented two different versions of the system, the "simple" one for old-hardware-robots like Wifibot-M along with some robot-specific implementation details, and the "complex" one for more efficient robots such as Aldebaran humanoid robots NAO and Pepper.

Perspectives

The combined visual attention model, that has been presented in this thesis, has met the objectives given in the General Introduction, and has proven itself to be applicable in real (or quasi-real) worlds tasks and environment. This being said, a number of perspectives remain still open and some aspects of the work could not be reasonably addressed in the limited timeframe of this thesis. These are subject for future development of the work accomplished here.

We can outline two different directions of perspectives, where first direction considers technical improvements, and the second is about further development of the model itself.

The nearest technical possibility of the system improvement, which could be predicted as a short-time perspective – the work of three or four months, – is the question of implementation. Both systems are implemented in Python, slow interpreted language with "global interpreter lock", which complicates parallel computations. Implementation translation into fast compilable languages, such as C++, would definitely improve the implementation performance.

Another point of technical improvement, more in long-term way, concerns the top-down mechanism. Both hardware and software are constantly evolving. The used approaches could be easily replaced with better analogues in the future, when and if these analogues will be created, – in several years from now.

The hardware improvement in the embedded CPU niche might also change the whole approach, if the whole combined system could be processed only by embedded hardware, thus pushing the robot further towards real autonomy. E.g., recent improvements in mobile CPU modules announce a support for Caffe or TensorFlow-based neural network implementations to be run directly on these mobile CPUs with a good level of efficiency[5]. Such an opportunity would mean a shift in priorities for the modules used in mechanisms such as ours, also moving the CNN processing to the embedded hardware. But still these CPUs and algorithms are a question of at least several years of development before they could be used in this field of research.

If speaking about further development of the model itself, we could say about further complexification of the attention model. As mid-term perspectives, – year or two, – we could talk about addition of several modules. E.g., Long-Term Memory in the line with Short-Term / Working Memory – representing more complex, real ontology constructions; or Audio Signal Treatment Module – in order to complete the phonological loop of the Working Memory, – these modules could grant higher

[5]More information on this could be found, for example, on the developer sites of Qualcomm: `https://developer.qualcomm.com/software/snapdragon-neural-processing-engine`

level cognition and autonomy.

We also should mention, that there already exists vast amount of knowledge stored in the Internet; thus, another different mid-term perspective would be the usage of third-party knowledge databases, e.g. BabelNet [Navigli 12] or WikiNet [Nastase 10] as semantic analysers in the decision module, or even as knowledge providers for the high-level decision making.

From another point of view onto the "Internet knowledge" for more long-term perspectives, Cloud Memory could become a very appealing mechanism, where several robots use the same knowledge storage through connection via Internet, providing also the learnt material into it, in order to "share the knowledge", thus creating some kind of Shared Long-Term Memory. Also, more accurate additional study into human psychoneurology could yield some additional bio-inspired mechanisms (e.g., information exchange techniques based on the human brain inter-cortex dynamic) to be in line with other human-brain-modelling concepts, such as "center-surround antagonism", "Baddeley-Hitch working memory" or "cognitive-phenomena-based attention".

Continuing the idea of "several robots" for long-term perspectives, we can also speculate on the idea of multi-agent distributed attention system, where each agent provides previously mentioned "shared knowledge". This could yield more accurate results for real-world problems, e.g. massive visual search or navigation networks.

Appendices

A | Things-50 Dataset

This dataset was produced by us using a mid-level camera, and consists of "easy-to-see" photos of different things, which could or could not be found in an office or in home. Each picture of such type contains one or several objects, of same or different types. The goal of usage of this dataset is to evaluate an approach which claims possibility to be able to detect the object's bounding rectangle and/or to recognize correctly (classify) the type of the object, contained inside; Figure A.1 contains several examples of this dataset.

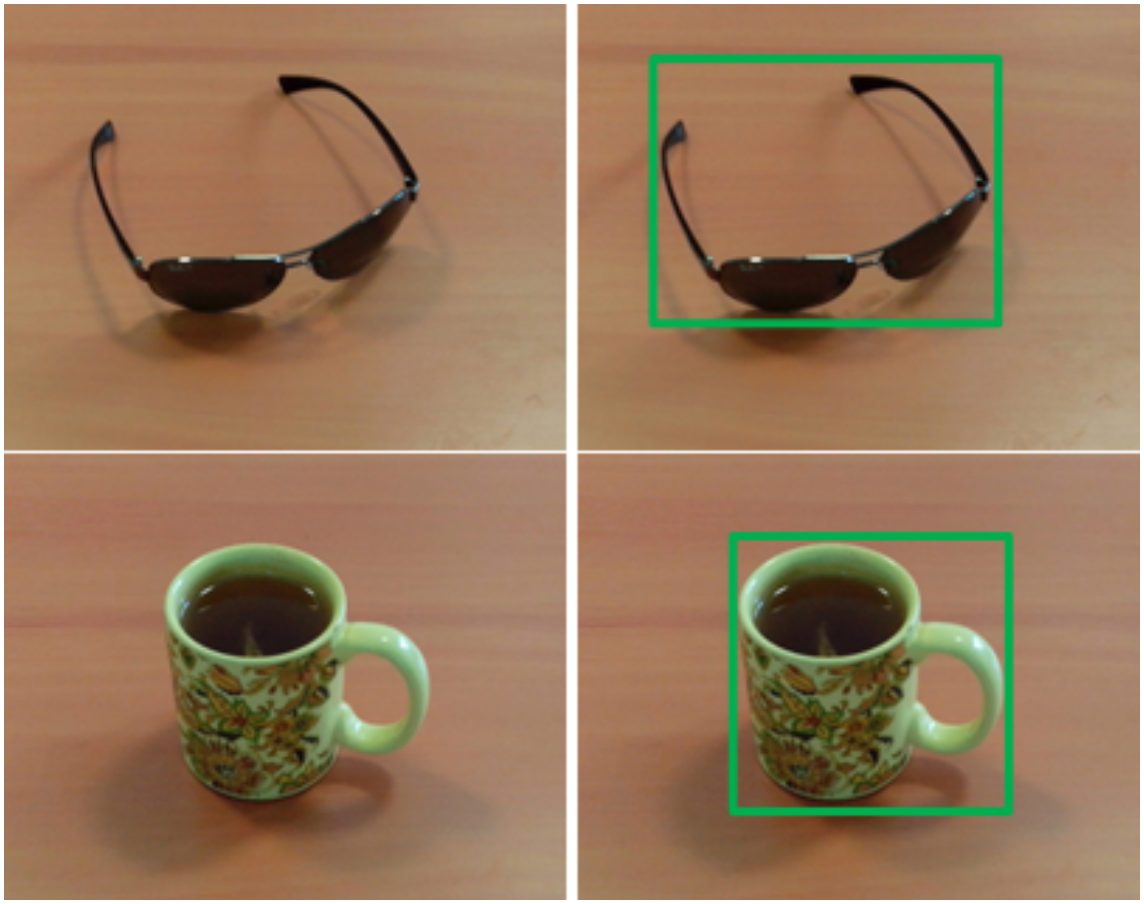

Figure A.1: Examples of images in self-created Things-50 database: sunglasses and a cup of tea, with ground-truth bounding rectangles

Auxiliary XML-files in this dataset contain also the markings of correct object classes – **n04356056** and **n07930864** (in WordNet notation [Miller 95], more on the subject is given in subsection 1.4.2), which can be decoded as synsets (syntactic sets) "sunglasses, dark glasses, shades" and "cup", respectively – the categories, used in ILSVRC2012 dataset (more on the subject is given in subsubsection 1.4.1.3).

An example of auxiliary XML is shown in Listing A.1.

```
<annotation>
        <folder>custom-lissi-images</folder>
        <filename>0009</filename>
        <source>
                <database>lissi-things-50</database>
        </source>
        <object>
                <name>n04356056</name>
                <top>110</top>
                <left>299</left>
                <height>183</height>
                <width>256</width>
        </object>
</annotation>
```

Listing A.1: Things-50 example XML of the sunglasses image

B | WiFiBot-M Controller Notes

This appendix chapter provides some notes on the software implementation for the WiFiBot-M robot management. While the Aldebaran humanoid robots have well-documented NAOqi API for simplification of their programming, the WiFiBot-M doesn't have any, as well as almost no documentation whatsoever and only some old-fashion proprietary software for its control, which is even difficult to be run on modern systems (programmed in C++ with usage of WinAPI and DirectConnect technologies, which makes this controller an old-Windows-only software).

The idea of creating new software controller comes from necessity of combining both movement controller and video stream controller in one software, which should be lightweight, and possible to be cross-platform.

Due to these conditions, a new controller is designed, "Wifibot-M iPy Controller", which should be able to:

- connect to robot via network, switch on/off its camera and/or wheels control;
- control movement of robot via Plug-and-Play joystick or its virtual simulation;
- collect video stream from camera, save stream frames as JPEG images;
- control camera's PTZ-routines;
- detect salient regions on video;
- react on these regions.

As a programming problem, application can be divided into 4 parts: movement controller, PTZ & video stream handler, salient regions detector and reaction strategies inspector. The general application structure was already presented in subsection 4.3.1 (Figure 4.3). To achieve cross-platform compatibility, while also being easy in terms of programming, a .NET platform (with its open source analogue Mono) has been chosen; programming language – IronPython.

Main module consists of GUI events and some routine work handling, which is why it is not shown on the conceptual scheme.

Second-main module of this application part is **WheelTalker** module, which implements connection between the robot itself and the whole application.

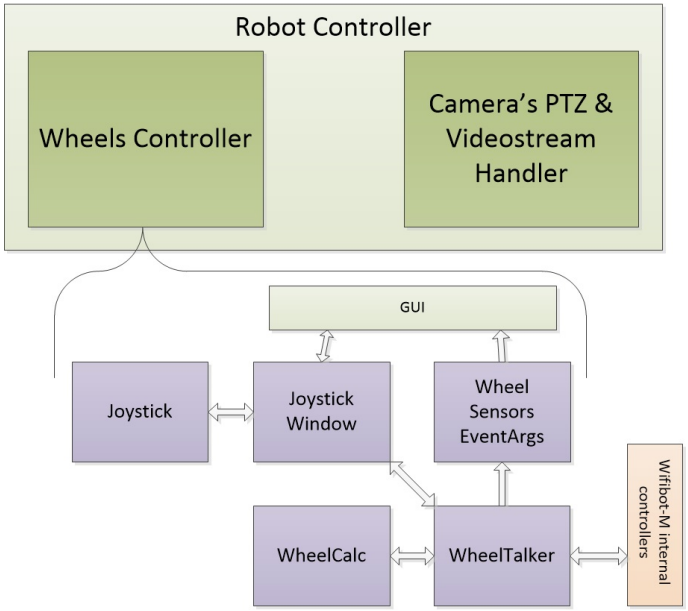

Figure B.1: Structure of Joystick and Wheels controller

Module **WheelTalker** is implemented as a separate thread, which has the following functionality implemented:

- network connection (TCP socket) between application and Wifibot-M internal controller;
- getting current sensor's data from internal controllers (speed, distance gone so far, battery level etc.);
- getting current joystick (real or virtual) position;
- calculating new TCP-package contents, which should be sent to Wifibot-M internal controllers to make it move properly.

Joystick position is taken from **JoystickWindow** module, new TCP-package contents is calculated in module **WheelCalc**, and sensors info is sent via **WheelSensorsEventArgs** event to GUI.

Module **WheelCalc** implements byte arithmetic over human-readable values, creating a 9-byte TCP-package payload, format of which is designed by robot's manufacturer. A code example of payload calculation is shown in Listing B.1. It uses Crc16 calculation routine, which is applied over payload bytes from 1 to 6.

```
pack[0] = 255
pack[1] = 0x07
tmp1 = 8*(speed1 & 0x3F)
tmp2 = 8*(speed2 & 0x3F)
pack[2] = tmp1 & 0xff
pack[3] = (tmp1 >> 8) & 0xff
pack[4] = tmp2 & 0xff
pack[5] = (tmp2 >> 8) & 0xff
if (speed2 & 0x80): tt=tt + 32
if (speed2 & 0x40): tt=tt + 16
pack[6] = (speed1 & 0x80) + (speed2 & 0x40) + tt + 13
mycrcsend = self.Crc16(pack, 1, 7)
pack[7] = mycrcsend & 0xff
pack[8] = (mycrcsend >> 8) & 0xff
```

Listing B.1: TCP initialization payload

Module **JoystickWindow** is implemented as a GUI region, where pointer offset describes current joystick's position. These offsets (X, Y) are treated as new speed values for left and right wheels – X is for direction change and Y is for forward-direction movement.

Module **Joystick** was initially created by Mark Harris[1], then rewrapped for IronPython usage. It uses DirectConnect module to interface a Plug-and-Play (usually through USB port) joystick. When real joystick is connected, this module reads its position and sends it as a new offset to **JoystickWindow** module.

The **Camera PTZ & Video stream handler** has two principal modules; the module which handles network connection and, therefore, PTZ movement and raw video data, and the module, which reforms the raw video data into a proper sequence of JPEG images (video frames).

First module, **CamTalker**, implements network connection (HTTP requests) to Wifibot camera's internal controller (AXIS Encoder module) through VAPIX protocol[2]. It executes requests and reads the responses, sending them to appropriate response handler (if needed).

Main handler of such responses is module **MjpegReader**, which has been made on base of Andrew Kirillov's implementation[3]. This module, working as a different

[1] http://airobots.googlecode.com/svn-history/r8/trunk/JoypadParser/JoystickInterface/Joystick.cs

[2] http://www.axis.com/techsup/cam_servers/dev/cam_http_api_index.php

[3] http://www.codeproject.com/Articles/15537/Camera-Vision-video-surveillance-on-C

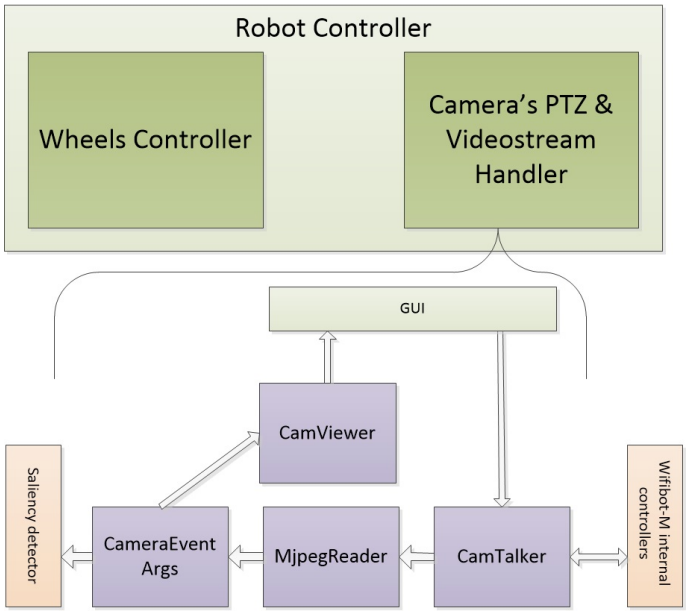

Figure B.2: Structure of Camera PTZ & Video stream handler

thread, reads whole HTTP-response, fetching single JPEG images as frames of video, sending them as events through **CameraEventArgs** to event handlers - saliency detector and a GUI video streaming.

Video streaming in GUI is made by **CamViewer** class, which is a basic GUI class with changed painting routines.

Publications

As the First Author

1. V. Kachurka, K. Madani, C. Sabourin, V. Golovko, “Visual Saliency Based Approach to Object Detection in Computer Vision Systems: Real Life Applications” in proc. of International Data Acquisition and Advanced Computing Systems (IEEE-IDAACS 2015), 20–24 September, 2015, Warsaw, Poland. – Vol.1, P. 174–182.

2. V. Kachurka, K. Madani, C. Sabourin, V. Golovko, “From Human Eye Fixation to Human-like Autonomous Artificial Vision” in proc. of International Work-Conference on Artificial Neural Networks as book “Advances in Computational Intelligence” (IWANN 2015), LNCS, Vol. 9094, Springer, Palma de Mallorca, Spain, 2015. – P. 171–184.

3. V. Kachurka, K. Madani, C. Sabourin, V. Golovko, “A statistical Approach to Human-Like Visual Attention and Saliency Detection for Robot Vision: Application to Wildland Fires’ Detection”, in Proc. of 8th International Conference on Neural Network and Artificial Intelligence (ICNNAI 2014), Communications in Computer and Information Science, Vol. 440, Springer, Brest, Belarus, 3–6 June 2014, pp. 124–135.

4. V. Kachurka, “Design Patterns in N-tier Architecture”, in Proc. of 15th International Ph.D. Workshop (OWD’2013), 19–22 October 2013, ISBN 978-83-935427-2-7, Wisla, Poland, pp. 94–97.

Journal Articles

1. K. Madani, V. Kachurka, C. Sabourin, V. Golovko, “A Human-like Visual-Attention-based Artificial Vision System for Wildland Firefighting Assistance", in Applied Intelligence, accepted, 2017.

2. K. Madani, V. Kachurka, C. Sabourin, V. Golovko, “A Soft-Computing-Based Approach to Artificial Visual Attention Using Human Eye-Fixation Paradigm: Toward a Human-Like Skill in Robot Vision", in Soft Computing, accepted, 2017.

Bibliography

[ABIResearch 13] ABIResearch. *Consumer and personal robotics.* `https://www.abiresearch.com/market-research/product/1014856-consumer-and-personal-robotics`, 2013. Accessed: 2016-10-14.

[Achanta 09] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada & Sabine Susstrunk. *Frequency-tuned salient region detection.* In Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on, pages 1597–1604. IEEE, 2009.

[Ackley 85] David H Ackley, Geoffrey E Hinton & Terrence J Sejnowski. *A learning algorithm for Boltzmann machines.* Cognitive science, vol. 9, no. 1, pages 147–169, 1985.

[Alahakoon 00] Damminda Alahakoon, Saman K Halgamuge & Bala Srinivasan. *Dynamic self-organizing maps with controlled growth for knowledge discovery.* IEEE Transactions on neural networks, vol. 11, no. 3, pages 601–614, 2000.

[Amarger 12] Véronique Amarger, Dominik Maximilián Ramík, Christophe Sabourin, Kurosh Madani, Ramón Moreno, Lucile Rossi & Manuel Graña. *Spherical coordinates framed RGB color space dichromatic reflection model based image segmentation: Application to wildland fires' outlines extraction.* In Image Processing Theory, Tools and Applications (IPTA), 2012 3rd International Conference on, pages 19–24. IEEE, 2012.

[Anderson 90] John R Anderson. Cognitive psychology and its implications. WH Freeman/Times Books/Henry Holt & Co, 1990.

[Atkinson 68] Richard C Atkinson & Richard M Shiffrin. *Human memory: A proposed system and its control processes.* Psychology of learning and motivation, vol. 2, pages 89–195, 1968.

[Baddeley 74] Alan D Baddeley & Graham Hitch. *Working memory.* Psychology of learning and motivation, vol. 8, pages 47–89, 1974.

[Baddeley 00] Alan Baddeley. *The episodic buffer: a new component of working memory?* Trends in cognitive sciences, vol. 4, no. 11, pages 417–423, 2000.

[Bailenson 05] Jeremy N Bailenson & Nick Yee. *Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments.* Psychological science, vol. 16, no. 10, pages 814–819, 2005.

[Ballard 07] Dana H Ballard, Mary M Hayhoe & Jeff B Pelz. *Memory representations in natural tasks.* Memory, vol. 7, no. 1, 2007.

[Bay 08] Herbert Bay, Andreas Ess, Tinne Tuytelaars & Luc Van Gool. *Speeded-up robust features (SURF).* Computer vision and image understanding, vol. 110, no. 3, pages 346–359, 2008.

[Behnel 11] Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn & Kurt Smith. *Cython: The best of both worlds.* Computing in Science & Engineering, vol. 13, no. 2, pages 31–39, 2011.

[Bekey 05] George A Bekey. Autonomous robots: from biological inspiration to implementation and control. MIT press, 2005.

[Benson 93] Scott Benson & Nils J Nilsson. *Reacting, Planning, and Learning in an Autonomous Agent.* In Machine intelligence 14, pages 29–64, 1993.

[Borji 10] Ali Borji, Majid Nili Ahmadabadi, Babak Nadjar Araabi & Mandana Hamidi. *Online learning of task-driven object-based visual attention control.* Image and Vision Computing, vol. 28, no. 7, pages 1130–1145, 2010.

[Borji 12a] Ali Borji. *Boosting bottom-up and top-down visual features for saliency estimation.* In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 438–445. IEEE, 2012.

[Borji 12b] Ali Borji & Laurent Itti. *Exploiting local and global patch rarities for saliency detection.* In Computer Vision and Pattern

Recognition (CVPR), 2012 IEEE Conference on, pages 478–485. IEEE, 2012.

[Borji 13a] Ali Borji & Laurent Itti. *State-of-the-art in visual attention modeling.* IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pages 185–207, 2013.

[Borji 13b] Ali Borji, Dicky N Sihite & Laurent Itti. *What stands out in a scene? A study of human explicit saliency judgment.* Vision research, vol. 91, pages 62–77, 2013.

[Borji 13c] Ali Borji, Hamed R Tavakoli, Dicky N Sihite & Laurent Itti. *Analysis of scores, datasets, and models in visual saliency prediction.* In Proceedings of the IEEE international conference on computer vision, pages 921–928, 2013.

[Borji 15] Ali Borji, Ming-Ming Cheng, Huaizu Jiang & Jia Li. *Salient object detection: A benchmark.* IEEE Transactions on Image Processing, vol. 24, no. 12, pages 5706–5722, 2015.

[Bradski 00] Gary Bradski. *The OpenCV Library.* Dr. Dobb's Journal: Software Tools for the Professional Programmer, vol. 25, no. 11, pages 120–123, 2000.

[Breazeal 99] Cynthia Breazeal & Brian Scassellati. *A context-dependent attention system for a social robot.* In International Joint Conference in Artifical Intelligence. Citeseer, 1999.

[Brownston 85] Lee Brownston, Robert Farrell, Elaine Kant & Nancy Martin. *Programming expert systems in OPS5.* 1985.

[Bruce 07] Neil Bruce & John Tsotsos. *Attention based on information maximization.* Journal of Vision, vol. 7, no. 9, pages 950–950, 2007.

[Buschman 07] Timothy J Buschman & Earl K Miller. *Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices.* science, vol. 315, no. 5820, pages 1860–1862, 2007.

[Bylinskii 16] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba & Frédo Durand. *What do different evaluation metrics tell*

us about saliency models? arXiv preprint arXiv:1604.03605, 2016.

[Bylinskii 17] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva & Antonio Torralba. *MIT Saliency Benchmark*, 2017.

[Cassagne 15] Ioannis Cassagne, Nicolas Riche, Marc Décombas, Matei Mancas, Bernard Gosselin, Thierry Dutoit & Robert Laganiere. *Video saliency based on rarity prediction: Hyperaptor.* In Signal Processing Conference (EUSIPCO), 2015 23rd European, pages 1521–1525. IEEE, 2015.

[Cerf 08] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser & Christof Koch. *Predicting human gaze using low-level saliency combined with face detection.* In Advances in neural information processing systems, pages 241–248, 2008.

[Chang 10] Chin-Kai Chang, Christian Siagian & Laurent Itti. *Mobile robot vision navigation & localization using gist and saliency.* In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pages 4147–4154. IEEE, 2010.

[Chen 09] Zhen-Xue Chen, Cheng-Yun Liu, Fa-Liang Chang & Guo-You Wang. *Automatic license-plate location and recognition based on feature salience.* IEEE transactions on vehicular technology, vol. 58, no. 7, pages 3781–3785, 2009.

[Cheng 15] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr & Shi-Min Hu. *Global contrast based salient region detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pages 569–582, 2015.

[Chum 03] Ondřej Chum, Jiří Matas & Josef Kittler. *Locally optimized RANSAC.* Pattern recognition, pages 236–243, 2003.

[Chun 98] Marvin M Chun & Yuhong Jiang. *Contextual cueing: Implicit learning and memory of visual context guides spatial attention.* Cognitive psychology, vol. 36, no. 1, pages 28–71, 1998.

[Clough 02] Bruce T Clough. *Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway.* Rapport technique, AIR

FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH, 2002.

[Comaniciu 02] Dorin Comaniciu & Peter Meer. *Mean shift: A robust approach toward feature space analysis.* IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 5, pages 603–619, 2002.

[Contreras-Reyes 12] Javier E Contreras-Reyes & Reinaldo B Arellano-Valle. *Kullback–Leibler divergence measure for multivariate skew-normal distributions.* Entropy, vol. 14, no. 9, pages 1606–1626, 2012.

[Courty 03] Nicolas Courty, Eric Marchand & Bruno Arnaldi. *A new application for saliency maps: Synthetic vision of autonomous actors.* In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, volume 3, pages III–1065. IEEE, 2003.

[Durkin 93] John Durkin. Expert systems: catalog of applications. Intelligent Computer Systems, 1993.

[Ebling 16] Maria R Ebling. *Can cognitive assistants disappear?* IEEE Pervasive Computing, vol. 15, no. 3, pages 4–6, 2016.

[Ehinger 09] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba & Aude Oliva. *Modelling search for people in 900 scenes: A combined source model of eye guidance.* Visual cognition, vol. 17, no. 6-7, pages 945–978, 2009.

[Einhäuser 08] Wolfgang Einhäuser, Merrielle Spain & Pietro Perona. *Objects predict fixations better than early saliency.* Journal of Vision, vol. 8, no. 14, pages 18–18, 2008.

[Erkan 04] Günes Erkan & Dragomir R Radev. *Lexrank: Graph-based lexical centrality as salience in text summarization.* Journal of Artificial Intelligence Research, vol. 22, pages 457–479, 2004.

[Fawcett 06] Tom Fawcett. *An introduction to ROC analysis.* Pattern recognition letters, vol. 27, no. 8, pages 861–874, 2006.

[Fraihat 15] Hossam Fraihat, Cristophe Sabourin & Kurosh Madani. *Soft-computing based fast visual objects' distance evaluation for*

robots' vision. In Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015 IEEE 8th International Conference on, volume 1, pages 81–86. IEEE, 2015.

[Francis 79] W Nelson Francis & Henry Kucera. *Brown corpus manual.* Brown University, vol. 2, 1979.

[Franke 05] Uwe Franke, Clemens Rabe, Hernán Badino & Stefan Gehrig. *6d-vision: Fusion of stereo and motion for robust environment perception.* In DAGM-Symposium, volume 3663, pages 216–223. Springer, 2005.

[Fritsch 08] J. Fritsch, T. Michalke, A. Gepperth, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko & C. Goerick. *Towards a human-like vision system for Driver Assistance.* In 2008 IEEE Intelligent Vehicles Symposium, pages 275–282, June 2008.

[Gates 07] B. Gates. *A robot in every home.* Scientific American, vol. 296, pages 58–65, 2007.

[Gavrilova 00] T.A. Gavrilova & V.F. Khoroshevskiy. *Knowledge bases of intellectual systems.* Piter, 2000. in Russian.

[Giovani 15] Bernardes Vitor Giovani, Alessandro Corrêa Victorino & Janito Vaqueiro Ferreira. *Stereo Vision for Dynamic Urban Environment Perception Using Semantic Context in Evidential Grid.* In Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, pages 2471–2476. IEEE, 2015.

[Goferman 12] Stas Goferman, Lihi Zelnik-Manor & Ayellet Tal. *Context-aware saliency detection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 10, pages 1915–1926, 2012.

[Golovko 03] Vladimir Golovko. *From neural networks to intelligent systems: Selected aspects of training, application and evolution.* NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES, vol. 186, pages 219–244, 2003.

[Gouk 14] Henry GR Gouk. Accelerating convolutional neural network systems. Master's thesis, The University of Waikato, 2014.

[Groover 07] Mikell P Groover. Fundamentals of modern manufacturing: materials processes, and systems. John Wiley & Sons, 2007.

[Hamm 10] Jordan P Hamm, Kara A Dyckman, Lauren E Ethridge, Jennifer E McDowell & Brett A Clementz. *Preparatory activations across a distributed cortical network determine production of express saccades in humans.* Journal of Neuroscience, vol. 30, no. 21, pages 7350–7357, 2010.

[Haring 94] S Haring, Max A Viergever & Joost N Kok. *Kohonen networks for multiscale image segmentation.* Image and vision computing, vol. 12, no. 6, pages 339–344, 1994.

[Hassan 15] Dayana Hassan, Kurosh Madani & Christophe Sabourin. *Dual 2-d images-based approach for objects'3-D characterization and localization for Machine-Awareness in indoor environment.* In Awareness Science and Technology (iCAST), 2015 IEEE 7th International Conference on, pages 201–206. IEEE, 2015.

[Hayes 08] Ian J Hayes. *Towards reasoning about teleo-reactive programs for robust real-time systems.* In Proceedings of the 2008 RISE/EFTS Joint International Workshop on Software Engineering for Resilient Systems, pages 87–94. ACM, 2008.

[Hayhoe 05] Mary Hayhoe & Dana Ballard. *Eye movements in natural behavior.* Trends in cognitive sciences, vol. 9, no. 4, pages 188–194, 2005.

[He 14] Kaiming He, Xiangyu Zhang, Shaoqing Ren & Jian Sun. *Spatial pyramid pooling in deep convolutional networks for visual recognition.* In European Conference on Computer Vision, pages 346–361. Springer, 2014.

[He 15] Kaiming He & Jian Sun. *Convolutional neural networks at constrained time cost.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5353–5360, 2015.

[Henderson 99] John M Henderson & Andrew Hollingworth. *High-level scene perception*. Annual review of psychology, vol. 50, no. 1, pages 243–271, 1999.

[Hering 74] Ewald Hering. *Zur Lehre von Lichtsinne. V. Grundz uge einer Theorie des Lichtsinnes*. SBK. Akad Wiss Wien Math natureiss K, vol. 69, pages 179–217, 1874.

[Hertz 91] John A Hertz, Anders S Krogh & Richard G Palmer. Introduction to the theory of neural computation, volume 1. Basic Books, 1991.

[Holland 92] John H Holland. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, 1992.

[Hopfield 82] John J Hopfield. *Neural networks and physical systems with emergent collective computational abilities*. Proceedings of the national academy of sciences, vol. 79, no. 8, pages 2554–2558, 1982.

[Huang 05] Jin Huang & Charles X Ling. *Using AUC and accuracy in evaluating learning algorithms*. IEEE Transactions on knowledge and Data Engineering, vol. 17, no. 3, pages 299–310, 2005.

[Hwang 11] Alex D Hwang, Hsueh-Cheng Wang & Marc Pomplun. *Semantic guidance of eye movements in real-world scenes*. Vision research, vol. 51, no. 10, pages 1192–1205, 2011.

[Imada 07] Akira Imada. *Finding a Needle in a Haystack: From Baldwin Effect to Quantum Computation*. In Computer Information Systems and Industrial Management Applications, 2007. CISIM'07. 6th International Conference on, pages 20–25. IEEE, 2007.

[Ittelson 76] William H Ittelson. *Environment perception and contemporary perceptual theory*. Environmental psychology: People and their physical settings, pages 141–154, 1976.

[Itti 98] Laurent Itti, Christof Koch & Ernst Niebur. *A model of saliency-based visual attention for rapid scene analysis*. IEEE

Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pages 1254–1259, 1998.

[Itti 07] L. Itti. *Visual salience: Scholarpedia article.* `http://www.scholarpedia.org/article/Visual_salience`, 2007. Accessed: 2015-10-20.

[Javal 78] Emile Javal. *Essai sur la physiologie de la lecture.* Annales d'Oculistique, vol. 80, pages 61–73, 1878.

[Jia 14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama & Trevor Darrell. *Caffe: Convolutional architecture for fast feature embedding.* In Proceedings of the 22nd ACM international conference on Multimedia, pages 675–678. ACM, 2014.

[Jiang 97] Jay J Jiang & David W Conrath. *Semantic similarity based on corpus statistics and lexical taxonomy.* arXiv preprint cmp-lg/9709008, 1997.

[Jiang 14] Ming Jiang, Juan Xu & Qi Zhao. *Saliency in crowd.* In European Conference on Computer Vision, pages 17–32. Springer, 2014.

[Joubert 08] Olivier R Joubert, Denis Fize, Guillaume A Rousselet & Michèle Fabre-Thorpe. *Early interference of context congruence on object processing in rapid visual categorization of natural scenes.* Journal of Vision, vol. 8, no. 13, pages 11–11, 2008.

[Judd 09] Tilke Judd, Krista Ehinger, Frédo Durand & Antonio Torralba. *Learning to predict where humans look.* In Computer Vision, 2009 IEEE 12th international conference on, pages 2106–2113. IEEE, 2009.

[Judd 12] Tilke Judd, Frédo Durand & Antonio Torralba. *A Benchmark of Computational Models of Saliency to Predict Human Fixations.* Rapport technique, MIT, 2012.

[Kachurka 12] Pavel Kachurka & Vladimir Golovko. *Fusion of recirculation neural networks for real-time network intrusion detection and*

recognition. International journal of computing, vol. 11, no. 4, pages 383–390, 2012.

[Kadir 01] Timor Kadir & Michael Brady. *Saliency, scale and image description*. International Journal of Computer Vision, vol. 45, no. 2, pages 83–105, 2001.

[Karatzas 15] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu*et al.* *ICDAR 2015 competition on robust reading*. In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pages 1156–1160. IEEE, 2015.

[Khan 14] Aslam Khan & Sanjay Mishra. *Image Compression using Growing Self Organizing Map Algorithm*. International Journal of Computer Science and Network Security (IJCSNS), vol. 14, no. 11, page 50, 2014.

[Khoroshevsky 93] Vladimir F Khoroshevsky. *Knowledge Based Design of Knowledge Based Systems in PiES WorkBench*. In Proc. Of Japan-CIS Symposium on Knowledge Based Software Engineering, volume 94, 1993.

[Kienzle 09] Wolf Kienzle, Matthias O Franz, Bernhard Schölkopf & Felix A Wichmann. *Center-surround patterns emerge as optimal predictors for human saccade targets*. Journal of vision, vol. 9, no. 5, pages 7–7, 2009.

[Klahr 87] David Klahr, Pat Langley & Robert Neches. Production system models of learning and development. MIT press, 1987.

[Koch 87] Christof Koch & Shimon Ullman. *Shifts in selective visual attention: towards the underlying neural circuitry*. In Matters of intelligence, pages 115–141. Springer, 1987.

[Koehler 14] Kathryn Koehler, Fei Guo, Sheng Zhang & Miguel P Eckstein. *What do saliency models predict?* Journal of vision, vol. 14, no. 3, pages 14–14, 2014.

[Koschate 16] Miriam Koschate, Richard Potter, Paul Bremner & Mark Levine. *Overcoming the uncanny valley: Displays of emotions reduce the uncanniness of humanlike robots.* In The Eleventh ACM/IEEE International Conference on Human Robot Interaction, pages 359–365. IEEE Press, 2016.

[Kosko 88] Bart Kosko. *Bidirectional associative memories.* IEEE Transactions on Systems, man, and Cybernetics, vol. 18, no. 1, pages 49–60, 1988.

[Krizhevsky 12] Alex Krizhevsky, Ilya Sutskever & Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks.* In Advances in neural information processing systems, pages 1097–1105, 2012.

[Kruthiventi 17] Srinivas SS Kruthiventi, Kumar Ayush & Radhakrishnan Venkatesh Babu. *Deepfix: A fully convolutional neural network for predicting human eye fixations.* IEEE Transactions on Image Processing, 2017.

[Kümmerer 16] Matthias Kümmerer, Thomas SA Wallis & Matthias Bethge. *DeepGaze II: Reading fixations from deep features trained on object recognition.* arXiv preprint arXiv:1610.01563, 2016.

[Leutenegger 11] Stefan Leutenegger, Margarita Chli & Roland Y Siegwart. *BRISK: Binary robust invariant scalable keypoints.* In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 2548–2555. IEEE, 2011.

[Li 10] Jia Li, Yonghong Tian, Tiejun Huang & Wen Gao. *Probabilistic multi-task learning for visual saliency estimation in video.* International journal of computer vision, vol. 90, no. 2, pages 150–165, 2010.

[Li 14] Jia Li, Yonghong Tian & Tiejun Huang. *Visual saliency with statistical priors.* International journal of computer vision, vol. 107, no. 3, pages 239–253, 2014.

[Lienhart 02] Rainer Lienhart & Axel Wernicke. *Localizing and segmenting text in images and videos.* IEEE Transactions on circuits and systems for video technology, vol. 12, no. 4, pages 256–268, 2002.

[Lin 13] Min Lin, Qiang Chen & Shuicheng Yan. *Network in network.* arXiv preprint arXiv:1312.4400, 2013.

[Liou 06] Cheng-Yuan Liou & Shiao-Lin Lin. *Finite memory loading in hairy neurons.* Natural Computing, vol. 5, no. 1, pages 15–42, 2006.

[Liu 11] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang & Heung-Yeung Shum. *Learning to detect a salient object.* IEEE Transactions on Pattern analysis and machine intelligence, vol. 33, no. 2, pages 353–367, 2011.

[Liu 16] Nian Liu & Junwei Han. *A deep spatial contextual long-term recurrent convolutional network for saliency detection.* arXiv preprint arXiv:1610.01708, 2016.

[Lowe 99] David G Lowe. *Object recognition from local scale-invariant features.* In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, volume 2, pages 1150–1157. Ieee, 1999.

[Lowe 04] David G Lowe. *Distinctive image features from scale-invariant keypoints.* International journal of computer vision, vol. 60, no. 2, pages 91–110, 2004.

[Lucas 03] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong & Robert Young. *ICDAR 2003 robust reading competitions.* In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, pages 682–687. IEEE, 2003.

[Mach 65] Ernst Mach. *Über die Wirkung der räumlichen Vertheilung des Lichtreizes auf die Netzhaut.* Sitzungsberichte der mathematisch-naturwissenschaftlichen Classe der kaiserlichen Akademic der Wissenschaftlen, vol. 52, pages 303–322, 1865.

[Madani 12] Kurosh Madani, Dominik M Ramik & Cristophe Sabourin. *Multilevel cognitive machine-learning-based concept for artificial awareness: application to humanoid robot awareness using visual saliency.* Applied Computational Intelligence and Soft Computing, vol. 2012, page 6, 2012.

[Mathur 16] Maya B Mathur & David B Reichling. *Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley.* Cognition, vol. 146, pages 22–32, 2016.

[Matsugu 03] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari & Yuji Kaneda. *Subject independent facial expression recognition with robust face detection using a convolutional neural network.* Neural Networks, vol. 16, no. 5, pages 555–559, 2003.

[Miksik 12] Ondrej Miksik & Krystian Mikolajczyk. *Evaluation of local detectors and descriptors for fast feature matching.* In Pattern Recognition (ICPR), 2012 21st International Conference on, pages 2681–2684. IEEE, 2012.

[Miller 95] George A Miller. *WordNet: a lexical database for English.* Communications of the ACM, vol. 38, no. 11, pages 39–41, 1995.

[Minsky 75] Marvin Minsky. *A framework for representing knowledge.* 1975.

[Moreno 11] Ramon Moreno, Manuel Grana, DM Ramik & Kurosh Madani. *Image segmentation by spherical coordinates.* In Proceedings of the 11th International Conference on Pattern Recognition and Information Processing (PRIP'11), pages 112–115, 2011.

[Nastase 10] Vivi Nastase, Michael Strube, Benjamin Börschinger, Cäcilia Zirn & Anas Elghafari. *WikiNet: A Very Large Scale Multi-Lingual Concept Network.* In LREC, 2010.

[Navalpakkam 05] Vidhya Navalpakkam & Laurent Itti. *Modeling the influence of task on attention.* Vision research, vol. 45, no. 2, pages 205–231, 2005.

[Navalpakkam 06] Vidhya Navalpakkam & Laurent Itti. *An integrated model of top-down and bottom-up attention for optimizing detection speed.* In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2049–2056. IEEE, 2006.

[Navalpakkam 10] Vidhya Navalpakkam, Christof Koch, Antonio Rangel & Pietro Perona. *Optimal reward harvesting in complex perceptual environments*. Proceedings of the National Academy of Sciences, vol. 107, no. 11, pages 5232–5237, 2010.

[Navigli 12] Roberto Navigli & Simone Paolo Ponzetto. *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*. Artificial Intelligence, vol. 193, pages 217–250, 2012.

[Nilsson 94] Nils Nilsson. *Teleo-reactive programs for agent control*. Journal of artificial intelligence research, 1994.

[Oliva 01] Aude Oliva & Antonio Torralba. *Modeling the shape of the scene: A holistic representation of the spatial envelope*. International journal of computer vision, vol. 42, no. 3, pages 145–175, 2001.

[Ouerhani 05] Nabil Ouerhani, Alexandre Bur & Heinz Hügli. *Visual attention-based robot self-localization*. In In Proceeding of European Conference on Mobile Robotics, pages 8–13, 2005.

[Ozuysal 10] Mustafa Ozuysal, Michael Calonder, Vincent Lepetit & Pascal Fua. *Fast keypoint recognition using random ferns*. IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 3, pages 448–461, 2010.

[Pan 17] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol & Xavier Giro-i Nieto. *SalGAN: Visual Saliency Prediction with Generative Adversarial Networks*. arXiv preprint arXiv:1701.01081, 2017.

[Parkhurst 02] Derrick Parkhurst, Klinton Law & Ernst Niebur. *Modeling the role of salience in the allocation of overt visual attention*. Vision research, vol. 42, no. 1, pages 107–123, 2002.

[Pedersen 04] Ted Pedersen, Siddharth Patwardhan & Jason Michelizzi. *WordNet:: Similarity: measuring the relatedness of concepts*. In Demonstration papers at HLT-NAACL 2004, pages 38–41. Association for Computational Linguistics, 2004.

[Peters 05] Robert J Peters, Asha Iyer, Laurent Itti & Christof Koch. *Components of bottom-up gaze allocation in natural images.* Vision research, vol. 45, no. 18, pages 2397–2416, 2005.

[Peters 07] Robert J Peters & Laurent Itti. *Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention.* In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007.

[Phillips 05] Joshua L Phillips & David C Noelle. *A biologically inspired working memory framework for robots.* In Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on, pages 599–604. IEEE, 2005.

[Pomplun 06] Marc Pomplun. *Saccadic selectivity in complex visual search displays.* Vision research, vol. 46, no. 12, pages 1886–1900, 2006.

[Powers 11] David Martin Powers. *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.* Bioinfo Publications, 2011.

[Ramík 13] Dominik Maximilián Ramík, Kurosh Madani & Christophe Sabourin. *From visual patterns to semantic description: A cognitive approach using artificial curiosity as the foundation.* Pattern Recognition Letters, vol. 34, no. 14, pages 1577–1588, 2013.

[Ramík 11] Dominik Maximilián Ramík, Christophe Sabourin, Kurosh Madani*et al. Hybrid salient object extraction approach with automatic estimation of visual attention scale.* In Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on, pages 438–445. IEEE, 2011.

[Ramík 12] Dominik Maximilián Ramík. *Contribution to complex visual information processing and autonomous knowledge extraction: application to autonomous robotics.* PhD thesis, Université Paris-Est, 2012.

[Rayner 98] Keith Rayner. *Eye movements in reading and information processing: 20 years of research.* Psychological bulletin, vol. 124, no. 3, page 372, 1998.

[Rensink 00] Ronald A Rensink. *The dynamic representation of scenes.* Visual cognition, vol. 7, no. 1-3, pages 17–42, 2000.

[Resnik 99] Philip Resnik*et al.* *Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language.* J. Artif. Intell. Res.(JAIR), vol. 11, pages 95–130, 1999.

[Riche 13a] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin & Thierry Dutoit. *Saliency and human fixations: State-of-the-art and study of comparison metrics.* In Proceedings of the IEEE international conference on computer vision, pages 1153–1160, 2013.

[Riche 13b] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin & Thierry Dutoit. *Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis.* Signal Processing: Image Communication, vol. 28, no. 6, pages 642–658, 2013.

[Rublee 11] Ethan Rublee, Vincent Rabaud, Kurt Konolige & Gary Bradski. *ORB: An efficient alternative to SIFT or SURF.* In Computer Vision (ICCV), 2011 IEEE international conference on, pages 2564–2571. IEEE, 2011.

[Russakovsky 15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein*et al.* *Imagenet large scale visual recognition challenge.* International Journal of Computer Vision, vol. 115, no. 3, pages 211–252, 2015.

[Schank 99] Roger C Schank. Dynamic memory revisited. Cambridge University Press, 1999.

[Scheier 97] Christian Scheier & Steffen Egner. *Visual attention in a mobile robot.* In Industrial Electronics, 1997. ISIE'97., Proceedings of the IEEE International Symposium on, volume 1, pages SS48–SS52. IEEE, 1997.

[Scragg 76] Greg Scragg. *Semantic nets as memory models.* Charniak and Wilks (1976), pages 101–127, 1976.

[Shell 00] Richard Shell. Handbook of industrial automation. CRC Press, 2000.

[Shen 14] Chengyao Shen & Qi Zhao. *Webpage saliency.* In European Conference on Computer Vision, pages 33–46. Springer, 2014.

[Simon 14] C. Simon. *Personal robotics: Market opportunities and Business models.* `http://innoecho.com/personal-robotics-market-opportunities-and-business-models-33`, 2014. Accessed: 2016-10-14.

[Simonyan 14] Karen Simonyan & Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.

[Skubic 04] Marjorie Skubic, David Noelle, Mitch Wilkes, Kazuhiko Kawamura & James M Keller. *A biologically inspired adaptive working memory for robots.* In AAAI Fall Symp., Workshop on the Intersection of Cognitive Science and Robotics: From Interfaces to Intelligence, 2004.

[Sowa 14] John F Sowa. Principles of semantic networks: Explorations in the representation of knowledge. Morgan Kaufmann, 2014.

[Sudol 10] Jeremi Sudol, Orang Dialameh, Chuck Blanchard & Tim Dorcey. *Looktel—A comprehensive platform for computer-aided visual assistance.* In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 73–80. IEEE, 2010.

[Szegedy 15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke & Andrew Rabinovich. *Going deeper with convolutions.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[Tatler 07] Benjamin W Tatler. *The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor*

biases and image feature distributions. Journal of vision, vol. 7, no. 14, pages 4–4, 2007.

[Tian 15] Yonghong Tian, Jia Li, Shui Yu & Tiejun Huang. *Learning complementary saliency priors for foreground object segmentation in complex scenes*. International Journal of Computer Vision, vol. 111, no. 2, pages 153–170, 2015.

[Toulouse 16] Tom Toulouse, Lucile Rossi, Turgay Celik & Moulay Akhloufi. *Automatic fire pixel detection using image processing: a comparative analysis of rule-based and machine learning-based methods*. Signal, Image and Video Processing, vol. 10, no. 4, pages 647–654, 2016.

[Treisman 80] Anne M Treisman & Garry Gelade. *A feature-integration theory of attention*. Cognitive psychology, vol. 12, no. 1, pages 97–136, 1980.

[Triesch 03] Jochen Triesch, Dana H Ballard, Mary M Hayhoe & Brian T Sullivan. *What you see is what you need*. Journal of vision, vol. 3, no. 1, pages 9–9, 2003.

[van de Weijer 04] Joost van de Weijer & Th Gevers. *Robust optical flow from photometric invariants*. In Image Processing, 2004. ICIP'04. 2004 International Conference on, volume 3, pages 1835–1838. IEEE, 2004.

[van Kleef 16] Joshua van Kleef. *Towards Human-like Performance Face Detection: A Convolutional Neural Network Approach*. Rapport technique, University of Twente, 2016.

[Vig 14] Eleonora Vig, Michael Dorr & David Cox. *Large-scale optimization of hierarchical features for saliency prediction in natural images*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2798–2805, 2014.

[Viola 04] Paul Viola & Michael J Jones. *Robust real-time face detection*. International journal of computer vision, vol. 57, no. 2, pages 137–154, 2004.

[Võ 12] Melissa L-H Võ, Tim J Smith, Parag K Mital & John M Henderson. *Do the eyes really have it? Dynamic allocation of attention when viewing moving faces.* Journal of vision, vol. 12, no. 13, pages 3–3, 2012.

[Walker 02] Laura L Walker & Jitendra Malik. *When is scene recognition just texture recognition.* Journal of Vision, vol. 2, no. 7, pages 255–255a, 2002.

[Wang 06] Jingdong Wang, Long Quan, Jian Sun, Xiaoou Tang & Heung-Yeung Shum. *Picture collage.* In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 1, pages 347–354. IEEE, 2006.

[Wang 12] Ting Wang, Dominik M Ramik, Christophe Sabourin & Kurosh Madani. *Intelligent systems for industrial robotics: application in logistic field.* Industrial Robot: An International Journal, vol. 39, no. 3, pages 251–259, 2012.

[Westheimer 04] Gerald Westheimer. *Center-surround antagonism in spatial vision: Retinal or cortical locus?* Vision research, vol. 44, no. 21, pages 2457–2465, 2004.

[Wolf 06] Christian Wolf & Jean-Michel Jolion. *Object count/area graphs for the evaluation of object detection and segmentation algorithms.* International Journal of Document Analysis and Recognition (IJDAR), vol. 8, no. 4, pages 280–296, 2006.

[Wolfe 05] Jeremy M Wolfe. *Guidance of visual search by preattentive information.* Neurobiology of attention, pages 101–104, 2005.

[Wolfe 07] Jeremy M Wolfe. *Guided search 4.0.* Integrated models of cognitive systems, pages 99–119, 2007.

[Wong 81] Eva Wong & Arien Mack. *Saccadic programming and perceived location.* Acta psychologica, vol. 48, no. 1, pages 123–131, 1981.

[Wu 94] Zhibiao Wu & Martha Palmer. *Verbs semantics and lexical selection.* In Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138. Association for Computational Linguistics, 1994.

[Yarbus 67] Alfred L Yarbus. *Eye movements during perception of complex objects*. In Eye movements and vision, pages 171–211. Springer, 1967.

[Zadeh 94] Lofti A Zadeh. *Fuzzy logic, neural networks, and soft computing*. Communications of the ACM, vol. 37, no. 3, pages 77–85, 1994.

[Zauner 10] Christoph Zauner. *Implementation and benchmarking of perceptual image hash functions*. 2010.

[Zelinsky 08] Gregory J Zelinsky. *A theory of eye movements during target acquisition*. Psychological review, vol. 115, no. 4, page 787, 2008.

[Zhang 08] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan & Garrison W Cottrell. *SUN: A Bayesian framework for saliency using natural statistics*. Journal of vision, vol. 8, no. 7, pages 32–32, 2008.

[Zhang 13] Jianming Zhang & Stan Sclaroff. *Saliency detection: A boolean map approach*. In Proceedings of the IEEE international conference on computer vision, pages 153–160, 2013.

[Zhou 17] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He & Jiajun Liang. *EAST: An Efficient and Accurate Scene Text Detector*. arXiv preprint arXiv:1704.03155, 2017.