Institut de Microtechnique
Université de Neuchâtel

Visual Attention: From Bio-Inspired Modeling to Real-Time Implementation

Nabil Ouerhani

thèse presentée à la faculté des sciences
pour l'obtention du grade de docteur ès sciences

Dissertation typeset with LaTeX 2_ε by the author.

IMPRIMATUR POUR LA THESE

Visual Attention : From Bio-Inspired Modeling to Real-Time Implementation

M. Nabil OUERHANI

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury

MM. H. Hügli (directeur de thèse),
P.-J. Erard, R. Müri (Berne),
et P.-Y. Burgi (Genève)

autorise l'impression de la présente thèse.

Neuchâtel, le 12 décembre 2003

La doyenne:

Martine Rahier

To my parents Hmida and Fatma

Abstract

Visual attention is the ability of a vision system, be it biological or artificial, to rapidly select the most salient and thus the most relevant data about the environment in which the system is operating. The main goal of this visual mechanism is to drastically reduce the amount of visual information that must be processed by high level and thus complex tasks, such as object recognition, which leads to a considerable speed up of the entire vision process.

This thesis copes with various aspects related to visual attention, ranging from biologically inspired computational modeling of this visual behavior to its real-time realization on dedicated hardware, and its successful application to solve real computer vision tasks. Indeed, the contributions brought together in this thesis can be divided into four main parts.

The first part deals with the computational modeling of visual attention by assessing the significance of novel features like depth and motion to the visual attention mechanism. Thereby, two models have been conceived and validated, namely the 3D- and the dynamic models of visual attention.

In the second part, the biological plausibility of various versions of the visual attention model is evaluated. Therefore, the performance of our visual attention model is compared with human visual attention behavior, assuming that the human visual attention is intimately linked to the eye movements.

The third part of the thesis covers our contribution on the realization of a real-time operating system of visual attention. Indeed, the computational model of visual attention is implemented on a highly parallel architecture conceived for general purpose image processing, which allows to reach real-time requirements.

Last but not least, the visual attention model has been successfully applied to speed up but also to increase the performance of various real tasks related to computer vision. Thereby, image compression, color image segmentation, visual object tracking, and automatic traffic sign detection and recognition largely benefit from the salient scene information provided by the proposed visual attention algorithm. Specifically, they use this information to automatically adjust their internal parameters according to scene contents, thus, considerably enhancing the quality of the achieved results.

⊢ ◊◊◊ ⊣

Résumé

L'attention visuelle est la capacité d'un système de vision, qu'il soit humain ou artificiel, de sélectionner, rapidement, les informations les plus pertinentes de l'environnement dans lequel il opère. Le rôle principal de ce mécanisme est de réduire sensiblement la quantité d'informations visuelles qui sera traitée par des tâches complexes telle que la reconnaissance d'objets, entraînant ainsi l'accélération de l'ensemble du processus de la vision.

Cette thèse couvre plusieurs aspects liés à l'attention visuelle, allant de la modélisation informatique bio-inspirée de ce mécanisme à sa réalisation en temps réel et son application pour résoudre des tâches pratiques liées à la vision par ordinateur. En fait, les contributions originales de cette thèse peuvent être divisées en quatre parties principales.

La première partie traite de la modélisation du mécanisme de l'attention visuelle en étudiant l'impact de nouvelles caractéristiques, comme la profondeur et le mouvement, sur le comportement du modèle informatique de l'attention. Suite à ces études, nous avons pu concevoir et implanter deux modèles d'attention visuelle, à savoir le modèle 3D et le modèle dynamique.

Dans la seconde partie de la thèse, nous évaluons la plausibilité biologique de quelques versions de notre modèle informatique d'attention visuelle. Pour ce faire, nous comparons la performance de notre modèle informatique au comportement de l'attention visuelle humaine, tout en supposant que celle-ci est étroitement liée aux mouvements oculaires.

La troisième partie expose notre contribution portant sur la réalisation d'un système d'attention visuelle opérant en temps réel. Il s'agit d'une implantation du modèle informatique d'attention sur une architecture de traitement d'image fortement parallèle ce qui répond parfaitement aux exigences temps-réel.

La dernière partie de cette thèse décrit comment le mécanisme de l'attention visuelle peut être appliquée à des tâches liées à la vision artificielle dans le but d'en accélérer le calcul mais aussi d'en augmenter la performance. La compression d'images, la segmentation d'images couleurs, le suivi d'objets et la détection et la reconnaissance automatique de panneaux routiers ont ainsi utilisé les informations pertinentes fournies par le mécanisme d'attention sur les scènes pour ajuster, d'une manière automatique, leurs paramètres internes en fonction du contenu de scènes, ce qui améliore nettement la qualité des résultats obtenus.

$\vdash \diamondsuit\diamondsuit\diamondsuit \dashv$

Kurzfassung

Visuelle Aufmerksamkeit ist die Fähigkeit eines biologischen oder künstlichen Sehsystems, schnell auffällige und daher relevante informationen einer Szene zu erfassen. Die Hauptrolle dieses Sehmechanismus ist es, die Menge der von höheren und deshalb komplexen kognitiven Modulen zu verarbeitenden Daten drastisch zu reduzieren, was den gesamten Sehprozess beträchtlich beschleunigt.

Die vorliegende Dissertation beschäftigt sich mit verschiedenen Aspekten der visuellen Aufmerksamkeit, sich erstreckend von der biologisch inspirierten Modellierung dieses Sehmechanismus bis zu seiner Echtzeit Implementierung auf dedizierter Hardware und seinem Einsatz im Bereich des Computer-Sehens, um praktische Anwendungen effizient zu unterstützen. In der Tat lassen sich die in der vorliegenden Arbeit geleisteten Beiträge in vier Hauptthemen unterteilen.

Der erste Teil behandelt die Modellierung der visuellen Aufmerksamkeit, indem die Auswirkungen neuartiger Merkmale wie Tiefe und Bewegung auf das Verhalten des Aufmerksamkeitsmechanismus abgeschätzt werden. Dadurch wurden zwei Computermodelle der Aufmerksamkeit konzipiert und implementiert, nämlich das 3D- und das dynamische Modell.

In dem zweiten Teil wird die biologische Plausibilität verschiedener Versionen unseres Aufmerksamkeitsmodelles untersucht. Dabei, vergleichen wir die Ergebnisse unseres Computermodelles mit dem menschlichen Aufmersamkeitsverhalten, unter der Annahme, dass die visuelle Aufmerksamkeit eines menschlichen Subjektes eng mit seinen Augenbewegungen verbunden ist.

Der dritte Teil der Dissertation beschreibt unseren Beitrag bezüglich der Entwicklung eines Echtzeit laufenden Aufmerksamkeitssystems. Da wurde unser Computermodell der visuellen Aufmerksamkeit auf eine hoch parallele Bildverarbeitungsarchitektur implementiert, was die Echtzeitanforderungen bestens erfüllt.

Letztlich, wird beschrieben, wie die visuelle Aufmerksamkeit erfolgreich eingesetzt wurde, um praktische Anwendungen im Gebiet des Computer-Sehens zu beschleunigen aber auch um ihre Leistungen zu erhöhen. Bildkompression, Segmentierung von Farbbildern, Objektverfolgung in Bildsequenzen und automatische Detektion und Erkennung von Strassenschildern ziehen dabei Vorteil aus den von der Aufmerksamkeitsmechanismus zu Verfügung gestellten relevanten Szene Informationen, um ihre internen Parameter auf den Inhalt der Szene anzupassen, was die Qualität der erzielten Ergebnisse eindeutig verbessert.

⊢ ◇◇◇ ⊣

Acknowlegments

I would like to express my gratitude to some persons who contributed directly or indirectly to the successful achievement of this thesis work.

First, my profound gratitude goes to Professor Heinz Hügli who gave me the very valuable opportunity to work in his research group, where I benefited from his rich scientific knowledge about vision, but also from the very agreeable working atmosphere that he created in the group thanks to his numerous human qualities.

I am deeply grateful to Professor Pieree-Jean Erard, Professor René Müri, and Dr. Pierre-Yves Buri, all of them members of the jury, for the time they spent to evaluate this dissertation and for their relevant feedbacks.

Some parts of the thesis work would not be achieved without numerous fruitful collaborations with other research institutions. I express my special thank to Dr. Friederich Heitger, Dr. Pierre-Yves Burgi, and Pierre-François Ruedi, members of the bio-inspired sensor division of the CSEM for providing their vision chip ProtoEye on which we realized the real-time visual attention system. The biological validation of our models were only possible thanks to the scientific knowledge and the facilities provided by the department of neurology at the university of Bern. The fruitful collaboration with Professor René Müri and Roman von Wartburg is highly appreciated. I had also the opportunity to initiate a collaboration regarding image segmentation with the department of computer science of our university. More precisely we had rich exchanges with Niculei Archip, then a member of the computer graphic group headed by Professor Pierre-Jean Erard. Last but not least, I am grateful to Dr. Javier Bracamonte, member of the signal processing group of our institute for the enriching collaboration regarding image compression.

I would like to thank Laurent Jeanrenaud and Heinz Buri, our current and former system managers. They were always available to efficiently solve the not so obvious computer problems. My special thank goes also to Claudine Faehndrich, Sandrine Piffaritti and to the whole administration staff for their help and kindness.

I am grateful to my colleagues of the pattern recognition group for the agreeable working atmosphere, but also for the fruitful scientific discussions. Timothee Jost, Olivier Huesser, and Thierry Zamofing were always attentive to my prob-

lems. I would like to express my special thank to Olivier Huesser for initiating me to numerous sport activities as soon as I arrived to Neuchâtel. I thank also all the colleagues from the signal processing group for the organization of numerous activities ranging from ski week-ends to *soirées pâtes*. Thanks also to all members of the IMTeam, our *victorious* soccer team, for their engagement and particularly for their persistency.

My special thanks go also to all students, from the university of Neuchâtel and EPF Lausanne, who carried out their Diploma and Semester projects under my supervision. In particular, I am grateful to Olivier Corbat who implemented the algorithms regarding detection and recognition of traffic signs.

Many friends from Neuchâtel, Basel and many other places in the world have supported me during these five years. Let them all find in these few lines my deep gratitude.

My special thank goes to Nuria for her continuous encouragement, technical help, and, particularly, for her moral support.

I would like to express my deep gratitude to my sisters and brothers: Mongia, Khaled, Fathia, Dalila, Zahida, Slim, and Aouatef who, despite the large distance separating us, provided me with enthusiastic encouragement and unconditional support.

Last but not least, I dedicate this thesis to my parents Hmida and Fatma for their generosity and affection. When I think about the sacrifices they made so that we could have a good education, I feel humble.

Nabil Ouerhani
Neuchâtel, November 2003

Contents

Abstract vii

Acknowledgments x

Table of Contents xiii

List of Figures xvi

List of Tables xix

List of Algorithms xx

List of Acronyms xxi

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Scope of the Thesis . . . 3
- 1.3 Main Contributions . . . 3
- 1.4 Thesis Outline . . . 3

2 State of the Art 5
- 2.1 Chapter Introduction . . . 5
  - 2.1.1 Chapter Outline . . . 5
- 2.2 Visual Attention in Humans . . . 5
  - 2.2.1 Anatomy of the Human Visual System . . . 6
  - 2.2.2 Organization of the Human Visual Attention System . . . 10
  - 2.2.3 Eye Movement and Visual Attention . . . 13
  - 2.2.4 Psychophysical Models of Visual Attention . . . 14
- 2.3 Visual Attention in Machines . . . 15
  - 2.3.1 Bio-inspired Computational Models of Attention . . . 15
  - 2.3.2 Visual Attention and Computer Vision Applications . . . 19
- 2.4 The Saliency-based Model of Visual Attention . . . 22
  - 2.4.1 Feature Maps . . . 22
  - 2.4.2 Conspicuity Maps . . . 24

2.4.3 Saliency Map . . . 25
2.4.4 Selection of Salient Locations . . . 26
2.4.5 Normalization Strategies for Map Combination . . . 26
2.5 Chapter Summary . . . 29

3 Extensions of the Basic Model of Visual Attention 31
3.1 Chapter Introduction . . . 31
3.1.1 Chapter Outline . . . 31
3.2 Model of Visual Attention for 3D Vision . . . 32
3.2.1 Depth from Stereo . . . 32
3.2.2 Conspicuity from Depth-Related Features . . . 34
3.2.3 Combining 2D- and 3D-Related Conspicuity Maps . . . 36
3.2.4 Results and Discussion . . . 37
3.3 Model of Dynamic Visual Attention . . . 42
3.3.1 Optical Flow Computation techniques . . . 42
3.3.2 Dynamic Conspicuity Map . . . 45
3.3.3 Combining Static and Dynamic Conspicuity Maps . . . 46
3.3.4 Results and Discussion . . . 49
3.4 Chapter Summary . . . 52

4 Empirical Validation of the Visual Attention Model 53
4.1 Chapter Introduction . . . 53
4.1.1 Chapter Outline . . . 54
4.2 Overview of the Method . . . 54
4.2.1 Computational Map of Attention . . . 54
4.2.2 Human Map of Attention . . . 55
4.2.3 Comparison Metrics . . . 57
4.3 Experiments and Discussion . . . 59
4.3.1 Validation of the $\mathcal{N}_1(.)$-based Model . . . 61
4.3.2 Validation of the $\mathcal{N}_2(.)$-based Model . . . 62
4.3.3 Color Contribution to Visual Attention . . . 63
4.3.4 Pop-out Effect . . . 66
4.3.5 Image Center Effect . . . 67
4.4 Chapter Summary . . . 68

5 Real-Time Visual Attention 69
5.1 Chapter Introduction . . . 69
5.1.1 Chapter Outline . . . 69
5.2 Vision Chips . . . 70
5.2.1 Fully Analog Vision Chips . . . 70
5.2.2 General Purpose Digital SIMD Architectures . . . 71
5.2.3 Mixed Analog-Digital SIMD Architectures . . . 71
5.3 ProtoEye: SIMD Machine for Image Processing . . . 72

5.3.1 Overview of the Architecture . . . 72
5.3.2 The Digital Part of ProtoEye . . . 73
5.3.3 The Analog Part of ProtoEye . . . 73
5.4 Implementation Issues . . . 76
5.4.1 Intensity Conspicuity Map . . . 76
5.4.2 Motion Conspicuity Map . . . 80
5.4.3 Saliency Map . . . 80
5.4.4 Detection of the Spots of Attention . . . 81
5.5 Experimental Results . . . 81
5.6 Performance Analysis and Perspectives . . . 83
5.6.1 Performance Analysis . . . 83
5.6.2 Perspectives . . . 86
5.7 Chapter Summary . . . 87

6 Application of Visual Attention to Computer Vision 89
6.1 Chapter Introduction . . . 89
6.1.1 Chapter Outline . . . 89
6.2 Focused Image Compression . . . 90
6.2.1 Baseline JPEG Algorithm . . . 90
6.2.2 Adaptive JPEG Algorithm . . . 91
6.2.3 Experimental Results . . . 92
6.3 Attentive Color Image Segmentation . . . 95
6.3.1 Spot-Based Color Image Segmentation . . . 96
6.3.2 MAPS: Multiscale Attention-based Pre-Segmentation of Color Images . . . 98
6.4 Attention-Based Object Tracking . . . 104
6.4.1 Object Detection and Characterization . . . 106
6.4.2 The Tracking Algorithm . . . 106
6.4.3 Perspectives . . . 108
6.5 Attention-Based Traffic Sign Recognition System . . . 110
6.5.1 Overview of the System . . . 111
6.5.2 Evaluation of the System . . . 112
6.6 Chapter Summary . . . 113

7 Conclusions 117

A Itti's Implementation 121
A.1 Multiscale Conspicuity Maps . . . 121
A.2 Gabor Pyramids . . . 121

B Gradient-Based Optical Flow 123
B.1 Gradient Constraint Equation . . . 123

References **125**

List of Figures

2.1 Structure of the human visual system. 6
2.2 Distribution of cones and rods in the retina. 7
2.3 Visual cortex and the *where* and *what* visual streams. 9
2.4 Structure of the human visual attention system I 11
2.5 Structure of the human visual attention system II 12
2.6 Example of a human scan path 13
2.7 Feature integration experiments. 14
2.8 Visual attention model as proposed by Koch and Ullman. 16
2.9 Selective Tuning model of attention 17
2.10 Attentive object recognition. 20
2.11 Overview of the NAVIS system. 21
2.12 Saliency-based model of visual attention. 23
2.13 Contents-based global amplification normalization. 27
2.14 Iterative non-linear normalization 28

3.1 Principle of stereoscopic vision. 32
3.2 Triclops. 33
3.3 Stereo image: Triclops. 34
3.4 Conspicuity from depth-related features 36
3.5 Results 1. 38
3.6 Results 2. 39
3.7 Results 3. 40
3.8 Results 4. 41
3.9 Large displacements and the multiscale concept 44
3.10 Computing of the dynamic conspicuity map 46
3.11 "Taxi Hamburg": dynamic conspicuity map. 47
3.12 Model of dynamic visual attention. 48
3.13 "Munich Train Station" sequence. 49
3.14 Integration of motion and color: competition-based strategy. . . . 51
3.15 Motion and color Integration: motion-conditioned strategy. 51
3.16 Tracking of the most salient moving objects. 52

4.1 Principle of eye movements recording. 55

4.2 Eye movement data example . . . 56
4.3 Fixation to saliency: Example . . . 58
4.4 Relative fixation-to-chance distance. . . . 59
4.5 Experiment images . . . 60
4.6 Impact of presentation time on Φ . . . 62
4.7 Φ: $\mathcal{N}_1(.)$-based model vs. $\mathcal{N}_2(.)$-based model . . . 64
4.8 ρ: $\mathcal{N}_1(.)$-based model vs. $\mathcal{N}_2(.)$-based model . . . 64
4.9 Color contribution to visual attention . . . 65
4.10 Color pop-out effect . . . 66
4.11 Image center effect. . . . 67

5.1 General purpose digital vision chip. . . . 72
5.2 ProtoEye platform. . . . 73
5.3 ProtoEye: single PE . . . 74
5.4 Characterization of the analog filter . . . 75
5.5 Variation of the diffusion length λ with V_G. . . . 76
5.6 $\mathcal{D}o\mathcal{E}xp$ versus $\mathcal{D}o\mathcal{G}$. . . . 78
5.7 ProtoEye resources allocation. . . . 79
5.8 Multiscale conspicuity transformation. . . . 81
5.9 Iterative normalization of conspicuity maps. . . . 82
5.10 Detecting the most salient locations . . . 82
5.11 Spots of attention from intensity and motion. . . . 84
5.12 Computation time: single map . . . 86

6.1 Baseline JPEG algorithm . . . 91
6.2 Adaptive JPEG algorithm: Quantizer. . . . 91
6.3 Quantizer of the adaptive JPPEG algorithm. . . . 92
6.4 Adaptive versus standard JPEG: Example 1. . . . 93
6.5 Adaptive versus standard JPEG: Example 2. . . . 94
6.6 Zoomed ROI from Example 2. . . . 95
6.7 Spot-based segmentation method . . . 97
6.8 MAPS: different steps . . . 98
6.9 MAPS: Segmentation-relevant data . . . 100
6.10 MAPS: Refinement procedure . . . 101
6.11 Computation of the typical value T of a presegmented region . . . 102
6.12 MAPS: Segmentation results . . . 103
6.13 Spot-based segmentation vs. MAPS segmentation. . . . 104
6.14 Attention-based object tracking . . . 105
6.15 Spot characterization by static and dynamic features . . . 107
6.16 Attention-based object tracking: An example . . . 110
6.17 Traffic sign models . . . 111
6.18 Attention-based traffic sign recognition system. . . . 113
6.19 Attention-based traffic sign recognition: Results . . . 114

List of Tables

4.1 Validation of the $\mathcal{N}_1(.)$-based Model 61
4.2 Validation of the $\mathcal{N}_2(.)$-based model 63

5.1 ProtoEye: Computation time of the complete process 85

6.1 Evaluation of the traffic sign detector. 114

List of Algorithms

4.1 Empirical method for model validation 54
4.2 From fixation points to human attention map 57
6.1 Seeded region growing algorithm 97
6.2 Attention-based object tracking 109

List of Acronyms

2D	Two Dimensional
3D	Three Dimensional
A/D	Analog to Digital converter
ALU	Arithmetic Logic Unit
CMOS	Complementary Metal-Oxide Semiconducter
CSEM	Centre Suisse d'Electronique et Microtechnique
DCT	Discrete Cosine Transform
CPU	Central Policy Unit
D/A	Digital to Analog converter
DMA	Dynamic Memory Access
$\mathcal{D}o\mathcal{G}$	Difference of Gaussians filter
$\mathcal{D}o\mathcal{E}xp$	Difference of Exponentials filter
$\mathcal{D}o\mathcal{O}r\mathcal{G}$	Difference of Oriented Gaussians filter
EMMA	Eye Movement and Movement of Attention
FEF	Frontal Eye Field
fMRI	functional Magnetic Resonance Imaging
FOA	Focus of Attention
FPGA	Field of Programmable Gate Array
IT	Inferior Temporal cortex
JPEG	Joint Photographic Experts Group
LGN	Lateral Geniculate Nucleus
LPF	Low Pass Filer
MAPS	Multiscale Attention-based PreSegmentation
MORSEL	Multiple Object Recognition and attentional SELection
MT	Middle Temporal area
NAVIS	Neural Active VISion
PC	Personal Computer
PE	Processing Element
PP	Posterior Parietal cortex
RF	Receptive Field
RGB	Red Green Blue
RISC	Reduced Instruction Set Computer
RN	Reticular Nucleus

ROI	Region Of Interest
SC	Superior Colliculus
SF	Scale Factor
SIMD	Single Instruction Multiple Data
SRG	Seeded Region Growing
VLSI	Very Large Scale Integration
WTA	Winner-Take-All

Chapter 1

Introduction

Vision is the most important of the five human senses, since it provides over 90 % of the information our brain receives from the external world. Its main goal is to interpret and to interact with the environments we are living in. In everyday life, humans are capable of perceiving thousands of objects, identifying hundreds of faces, recognizing numerous traffic signs, and appreciating beauty almost effortlessly. The ease with which humans achieve these tasks is in no way due to the simplicity of the tasks but is a proof of the high degree of development of our vision system.

Computer vision is an applied science whose main objective is to provide computers with the functions present in human vision. Typical applications of computer vision are robot navigation, video surveillance, medical imaging, industrial quality control, to mention only some of them. Despite the impressive progress made in this field during the last decades, the currently available computer vision solutions by far underlay the human visual system regarding robustness and performance.

Computer vision systems that are inspired from human vision represent a promising alternative towards building more robust and more powerful computer vision solutions. To develop such systems, researchers might first answer the question "what are the mechanisms involved in vision that make it apparently so easy for humans, yet so difficult for the computer?"

Visual attention refers to the ability of vision systems to rapidly select the most salient and thus the most relevant data in a scene. The main goal of this visual behavior is to drastically reduce the amount of visual information that must be processed by high level and thus complex tasks, such as object recognition, leading, thereby, to a considerable speed up of the entire vision process. We believe that this visual mechanism represents a non-negligible part of the answer to the question asked above.

1.1 Motivation

It is generally admitted [95] that human vision can basically be divided into two main phases, low-level vision and high-level vision. Although the frontier between the two vision phases is not clearly defined, their main roles have already been established.

Low-level vision starts with gathering visual information by means of the retinas in the order of 10^8 bits/second [69]. The gathered information is then transmitted to the visual cortex where information about motion, depth, color, orientation, and shape is extracted. The high-level vision performs then its task on these extracted features. It is mainly responsible for recognizing scene contents by matching the representative scene features to a huge database of learned and memorized objects (i.e. object recognition).

Despite the huge amount of visual information to be processed and despite the combinatorial nature of the recognition task, it has been estimated that humans can recognize a large number of objects in less than 200 ms [144, 145]. The computation resources of the human brain can not entirely explain this astonishing performance, since the about 10^{11} neurons of our brain can not process such amount of information in such a short time, given their slow response rate. This high performance speaks rather for the high efficiency of our vision system.

The existence of a mechanism that selects only a reduced set of the available visual information for high level processing seems to be the most coherent explanation of the high performance of the human visual system [150]. This mechanism is referred to as visual attention. The hypothesis of the existence of a visual attention mechanism is reinforced by the anatomical structure of the retina itself. In fact, the inhomogeneous distribution of the photoreceptors over the retina leads to a precise sensing of only a reduced part of the visual field, whereas the rest part is only vaguely perceived. Hence, in order to perceive the most informative parts of the visual field a shift of the central part of the retina (i.e. fovea) is necessary. The visual attention mechanism seems to control the selection of the informative parts of the scene to which the fovea is oriented.

In computer vision, the computational complexity of numerous tasks like perceptual grouping and object recognition, which are known to be NP-Complete, represents a fundamental obstacle for real-world applications. Thus, the visual attention paradigm is a highly relevant topic if we want to master the complexity issue in computer vision [150, 151]. Indeed, visual attention can be conceived as a preprocessing step which permits a rapid selection of a subset of the available sensory information. Once detected, the salient parts become the specific scene locations on which higher level computer vision tasks can focus.

1.2 Scope of the Thesis

The present thesis deals with visual attention. Thanks to the latest findings about the human visual attention on one hand and the fast performance growth of computers on the other, it becomes possible to simulate many aspects of this visual behavior. In addition, the emergence of interdisciplinary approaches to vision creates a fruitful interaction between scientists of different fields, which leads to the conception of bio-inspired artificial vision systems. Furthermore, the increasing demand for real-time computer vision solutions for problems of increasing complexities encourages the consideration of the visual attention paradigm in computer vision tasks.

Indeed, this thesis copes with different aspects of visual attention ranging from biologically inspired modeling of this visual behavior to its real-time implementation on an artificial vision system. The application of the visual attention paradigm in real-world computer vision tasks represents also a central part of this thesis, since basic tasks like image segmentation, but also more advanced ones, such as object tracking and traffic sign recognition, have largely benefited from the visual attention mechanism.

1.3 Main Contributions

This thesis extensively relies on previous works that have reported impressive findings on biological and artificial visual attention. The saliency-based model of visual attention presented in [79, 74, 68] has been the start point of our work. The following points have not, however, been covered by previous works and thus represent the novel contributions of the present thesis.

- Extension of the saliency-based model of visual attention to operate on 3D scenes on one hand [114] and on dynamic scenes on the other hand [116].
- Empirical validation of the saliency-based model of visual attention by comparing its performance with the human visual attention behavior [157, 119].
- Real-time implementation of the saliency-based model of visual attention on a highly parallel, low-power, single board SIMD architecture [118, 117].
- Novel application of the visual attention algorithm in various fields like image compression [112], image segmentation [110, 111, 115], object tracking [116], and traffic sign recognition [34].

1.4 Thesis Outline

The remainder of this thesis is structured into six chapters. Chapter 2 deals with the state of the art in the context of visual attention. It describes some works

that contributed to a better understanding of the attention mechanism, be it in biology, psychology or computer science. It also gives a detailed description of the saliency-based model of visual attention, on which the present work has been based.

Chapter 3 reports two major extensions of the basic saliency-based model of visual attention. The first extension consists in integrating depth (3D) scene features into the basic model, which considers only 2D-based features. The second extension gives rise to the model of dynamic visual attention by considering also dynamic scene features as source of attention.

Chapter 4 deals with the empirical validation of the saliency-based model of visual attention. Comparing the results provided by the computational model with the human attention, the validation method quantitatively evaluates the contribution of color to visual attention controlling. Also, it assesses the plausibility of different working modes of the model, such as the linear and non-linear combination of visual cues into the final attention map.

A real-time implementation of the saliency-based model of visual attention is reported in Chapter 5. This implementation is achieved on a Single Instruction Multiple Data (SIMD) architecture that has been conceived for general purpose low-level image processing.

Chapter 6 presents how visual attention can be considered to efficiently guide some tasks related to computer vision. Four different applications are considered in this context: adaptive image compression, color image segmentation, object tracking, and automatic traffic sign recognition for intelligent vehicles.

Finally, Chapter 7 concludes this thesis and gives challenging perspectives of our work.

Chapter 2

State of the Art

2.1 Chapter Introduction

Visual attention has been an important investigation field for the last few decades. Be it in neurology, psychology or even in computer science, the scientific community has made impressive advances towards understanding and modeling this visual behavior. The emergence of interdisciplinary researches has yielded a mutual benefit for both human and machine vision communities. This chapter highlights some of the scientific findings that permitted a better understanding of visual attention and its successful application in computer vision.

2.1.1 Chapter Outline

Section 2.2 of the current chapter reports biological as well as psychophysical contributions on visual attention. Section 2.3 deals with artificial visual attention. It reports some computational models of visual attention known from the literature and describes how these models have been used to solve some issues in computer vision applications. Section 2.4 gives a detailed description of one of the most used computational models of visual attention, namely the saliency-based model.

2.2 Visual Attention in Humans

Visual attention is basically a biological mechanism used essentially by primates to compensate for the inability of their brains to process the huge amount of visual information gathered by the two eyes. This section reports some anatomical proofs about the existence of a visual attention mechanism present in the human vision and, later on, describes some psychophysical studies about it, which allows the modeling of the human attention behavior.

2.2.1 Anatomy of the Human Visual System

This section gives a general overview of the components of the human visual system. Far from being a detailed description of the human vision system, this overview helps the reader to understand why we, humans, need visual attention.

To learn more about some physiological and neurological aspects of human vision, the reader is referred to a wealth of publications on this subject [86, 18, 134].

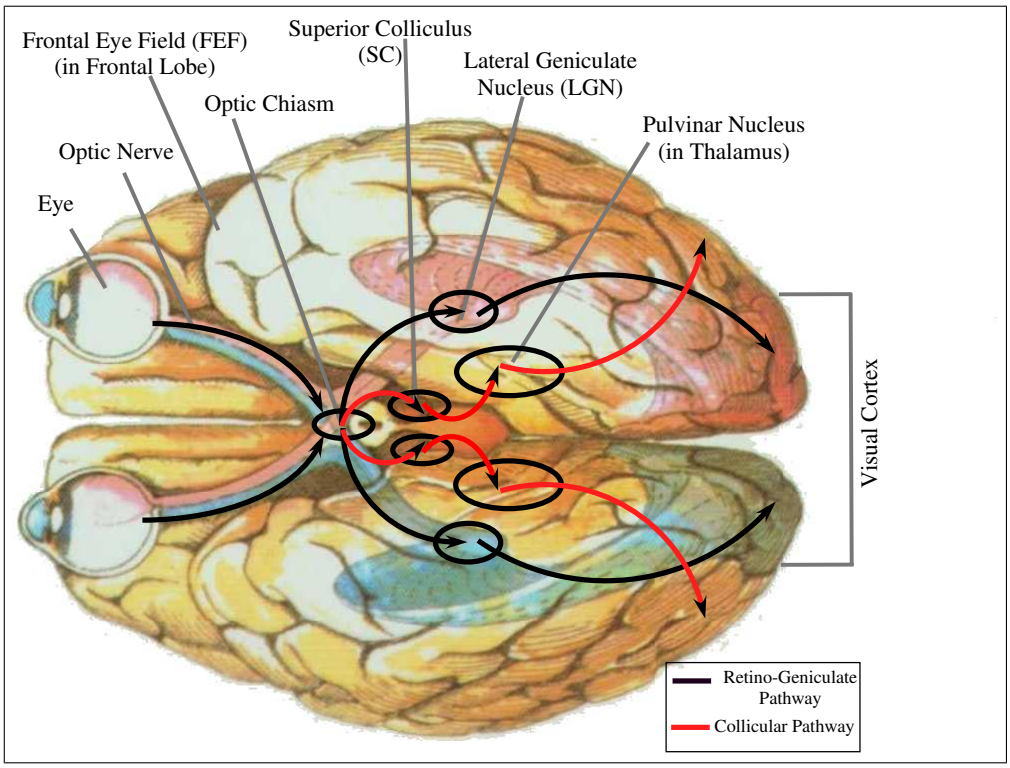


Figure 2.1: Structure of the human visual system (adapted from [106]).

Overview

The visual information collected by the retina is transmitted by the optic nerve to two different brain centers; the lateral geniculate nucleus (LGN), which is part of the thalamus, and the superior colliculus (SC). The flow of the visual information into various brain areas gives rise to two main pathways, namely the retino-geniculate and the collicular. The retino-geniculate pathway transmits about 90% of the visual information into the visual cortex, involved with spatial, temporal, chromatic, and disparity processing. Since the collicular pathway is responsible for only 10% of the entire visual information, it is often neglected

when describing in general terms the human visual system. It will be, however, considered in this thesis, because of its important role in visual attention and eye movements [42].

Figure 2.1 depicts the main components of the human visual system as well as the main visual information flows from the retina to the visual cortex.

The Retina

The retina forms the inside lining of the back of the eye and is composed of photosensitive cells, which can be classified into two categories: rods and cones. The retina contains about $5 \cdot 10^6$ cones responsible for photopic and color vision. The rods, whose number amounts to about 10^8, are essential for the nocturnal vision, and they are generally neglected in studies of visual acuity [18]. They will not be further considered in this report.

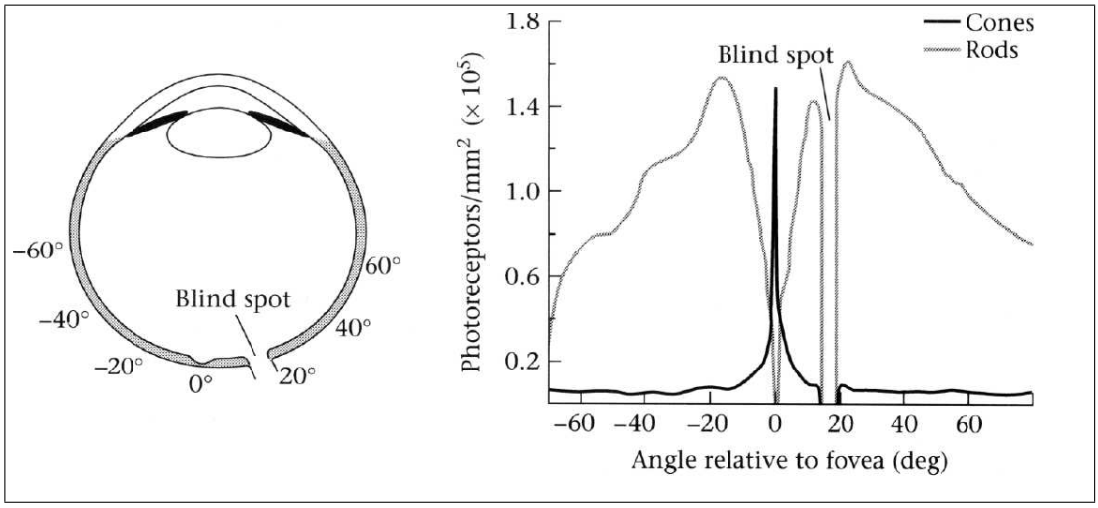


Figure 2.2: Distribution of cones and rods in the retina (from [134]).

The cones are not uniformly distributed in the retina. A high concentration of cones is found centrally in a small part named the fovea. In the more peripheral part, the concentration of cones decreases with the eccentricity. Due to this non-uniformity in cones distribution, high resolution visual information is available only within a reduced part of the visual field. Figure 2.2 graphically illustrates the spatial distribution of the photoreceptors over the retina.

The retinal photoreceptors are connected (via bipolar cells) with the ganglion cells. These cells can be classified into three main categories: the M (for magno) ganglion cells, the P (for Parvo) ganglion cells and a third category which englobes the non-M non-P cells. The P ganglion cells mostly connect with foveal cones and thus are associated with central vision, whereas M ganglion cells mostly connect with cones in the periphery.

In addition to their different distribution over the retina, the P- and the M-cells differ from each other regarding characteristics like spatial and temporal resolution. On one hand, P-cells have smaller receptive fields (RF) than M-cells and thus higher spatial resolution. On the other hand, the M-cells have higher temporal resolution, since they best respond to transient patterns, whereas the P-cells give sustained response only to stationary stimuli.

We will see later that these cell types give rise to different visual streams.

The Thalamus

The thalamus is located in the anterior and dorsal side of the midbrain and is composed of a number of nuclei, namely the lateral geniculate nucleus (LGN), the pulvinar nucleus, and the reticular nucleus (RN).

The lateral geniculate nucleus forms the main relay of visual information to the cerebral cortex. It is a cortical area that segregates the various retinal subsystems serving the contralateral visual fields and organizes their projections to the visual cortex via the retino-geniculate pathway. The cells constituting the LGN are of three different types: *parvocellular*, *magnocellular* and *koniocellular*. The parvocellular and the magnocellular cells are synapsed respectively with the P and M ganglion cells of the retina. It is generally admitted that these two types of cells feeds two different visual streams of the retino-geniculate pathway: the parvocellular and the magnocellular visual streams [86]. The non-M non-P ganglion cells of the retina feed into the koniocellular cells of the LGN. Recent works [64] proved the existence of this third stream related to this type of cells, and forming the koniocellular stream.

The pulvinar is the largest nucleus of the thalamus, since it occupies the two-fifth of its entire volume. Although it has no direct input from the optic nerve, the pulvinar has reciprocal connections with all cortical areas. An important role in attention guiding and eye movements has been assigned to this nucleus [84, 83].

Another thalamic nucleus which is believed to influence the attention mechanism is the reticular nucleus (RN). The cells of this nucleus seem to be modulated by an alerting-related signal [16, 96], that is a signal generated during an alert state in order to accelerate the selection of the visual information and thus to allow for instance rapid reaction to danger.

The Visual Cortex

The about 10^{10} cells of the visual cortex are organized in a hierarchical manner. The area V1, also called striate cortex is at the basis of the hierarchy, since it represents the major entry of the visual information coming from the LGN to the visual cortex. It is noteworthy that 50% of the area of V1 is devoted to the

representation of the visual information gathered by the central part of the retina (the fovea).

From V1, the visual information is then transmitted to the area V2. Thereafter, the two visual pathways seem to diverge, since the parvocellular stream feeds to V4, whereas the magnocellular one feeds to Middle Temporal area (MT). V4 is associated to the synthesis of complex forms and to combination of color and form. It transmits the processed visual information, via the parvocellular stream, to inferior temporal cortex (IT), which represents the final visual area for object recognition. For this reason, the parvocellular stream is also known as the *what* stream.

The middle temporal (MT) area represents an important route of visual information to the posterior parietal cortex (PP), which is responsible for object localization. Thus, the magnocellular pathway is often referred to as the *where* stream. The *where* stream is believed to play a fundamental role in controlling visual attention deployment [73].

Figure 2.3 schematizes the visual information flow between the visual cortex areas.

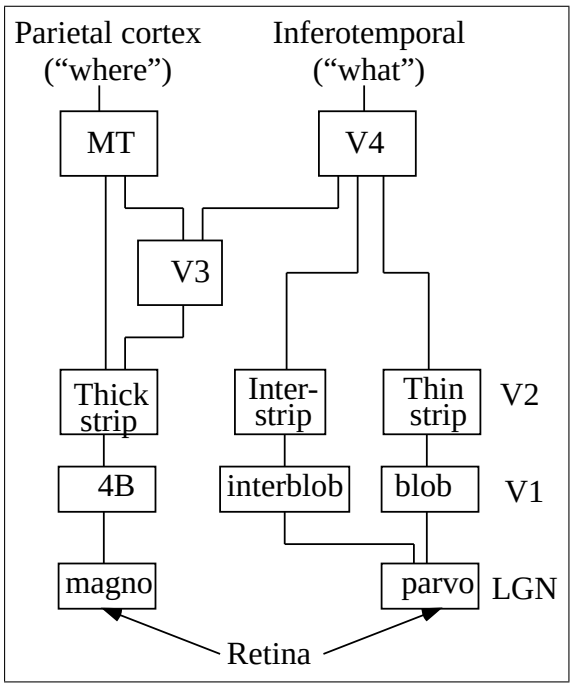


Figure 2.3: Visual cortex and the *where* and *what* visual streams (from [155]).

Superior Colliculus and Frontal Eye Field

This section describes two additional visual brain centers, whose role in controlling visual attention and eye movements has been established, namely the superior colliculus (SC) and the frontal eye field (FEF).

The superior colliculus is located on the dorsal surface of the brain stem (see Figure 2.1). It receives visual inputs directly from the retina and also from the visual cortex. The retinal inputs are mainly restricted to M ganglion cells. Superior colliculus projects to numerous nuclei of the thalamus which in turn project to cortical areas involved with saccadic eye movements [78]. Various hypotheses suggest that the SC strongly contributes to the building of the global map of attention (saliency map) [42, 7].

The frontal eye field which is located in the frontal lobe of the brain (see Figure 2.1) is involved in the generation of ocular movements. It receives input from different areas of the visual cortex and is directly connected to the oculomotor centers. Some works have suspected the FEF to harbor attention controlling mechanisms [143, 61].

To summarize, the human visual system is organized so that only small portion of the visual field can be perceived in high resolution, namely by the central part of the retina - the fovea -. Furthermore, the major part of the processing capacity of the visual cortex is devoted to the foveal visual information. Thus, to analyze the entire visual field, a successive deployment of the fovea to cover the important parts of the scene is necessary.

Note that the reported findings about the existence of a visual attention mechanism are not limited to humans but are also valid for most primates.

2.2.2 Organization of the Human Visual Attention System

The previous section pointed out the implication of some brain centers in the visual attention control. However, a clear definition of their respective role is still not possible. Nevertheless, the knowledge available about the human visual attention system, gained either through functional Magnetic Resonance Imaging (fMRI) or through patients examination allowed to establish some hypotheses about the organization of the visual attention system. In the following, we briefly describe three of these hypotheses which all agree that the visual attention system is distributed over a network of brain areas. For a more complete overview of existing hypotheses about visual attention organization, the reader is referred to [56].

Posner's Hypothesis

According to Posner [124, 123], attention in general and visual attention in particular sequentially involves the following three operations: first, attention must be disengaged from its current locus; next it must be moved to the new target; and finally it must be engaged on the new selected location. Posner suggested that the disengage operation is controlled by the posterior parietal cortex (PP), whereas the move step is controlled by the midbrain, in particular by the superior colliculus (SC). The engage operation depends on the pulvinar (a thalamic nucleus). The three attention operations as well as their respective brain areas are illustrated in Figure 2.4(a).

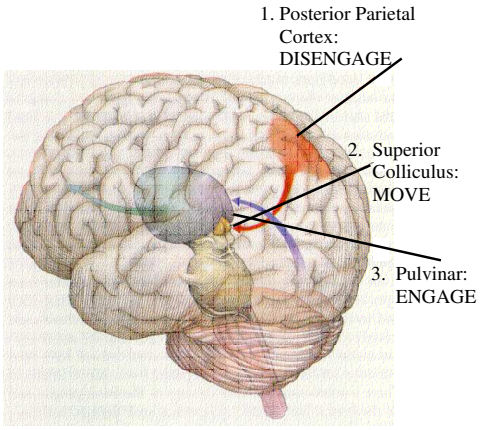
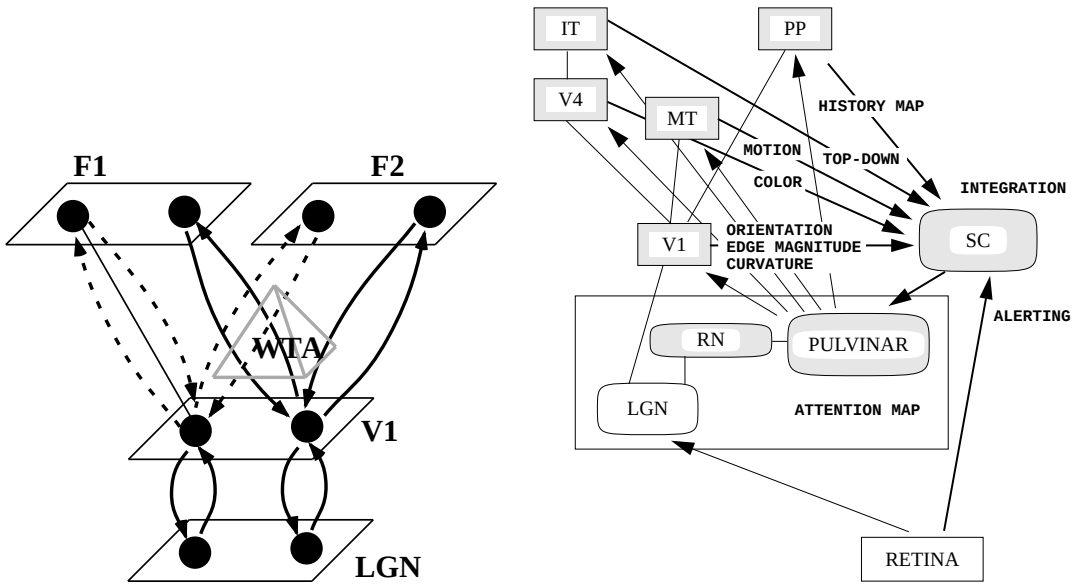


Figure 2.4: Structure of the human visual attention system: Posner's hypothesis (after [124]).

Caltech's Hypothesis

The hypothesis elaborated by Koch and Ullman in [79] and reinforced by Niebur in [104, 105] represents one of the first concrete descriptions of the cortical areas involved in controlling the visual attention mechanism. The authors went further to conceive a computational model that matches their hypothesis, namely the saliency-based model of visual attention (a detailed description of this model is given in Section 2.3). According to this hypothesis, elementary features are extracted in cortical area within and beyond the striate cortex V1, for instance MT for motion and V4 for color. The authors suggested that these features are combined in a unique map of attention, the saliency map, which resides either in LGN or in V1. Finally, the Winner-Take-All (WTA) network which is responsible for detecting the most salient scene location is suspected to be harbored by the thalamic reticular nucleus. A schematic description of this hypothesis is given in Figure 2.5(a).



(a) Caltech's hypothesis (b) Milanese's Hypothesis

Figure 2.5: Structure of the human visual attention system: (a) Caltech's hypothesis (from [79]); (b) Milanese's hypothesis (from [96]).

Milanese's Hypothesis

Taking advantage of the work of Posner [124] and of that of Desimone [42], Milanese has extended the Caltech's hypothesis to a more detailed description of the human visual attention system [96] (Figure 2.5(b)). Like Koch, Milanese suggested that the cortical areas are responsible for computing feature maps (color, orientation, motion, ...), which are integrated into a final map of attention. The superior colliculus is suggested to orchestrate the integration process, whereas the thalamus (in particular the pulvinar and the reticular nucleus) is supposed to harbor the final map of attention. Top-down or task dependent influences on the attention behavior seems to be received by the SC from the inferior temporal cortex (IT), whereas alert signals can be received by the same brain area directly from M-cells of the retina. The parietal cortex (PP) which has the capability to assess spatial relations is suggested to influence the selection of new targets on the basis of their spatial relation with previously attended locations.

The authors of the three hypotheses pointed out that most of the brain areas which control the visual attention mechanism are also involved in the control of eye movement. Given the anatomical link between attention and eye movement, next section is devoted to the description of psychophysical findings which reveal the relationship between the two behaviors.

2.2.3 Eye Movement and Visual Attention

There exist numerous types of eye movement [75]. The most studied ones are the saccadic eye movements [53], which are responsible for shifting the high-resolution fovea onto a given target. Once foveated, this target can be processed with more details. Thus, while exploring a given scene humans shift, successively, their fovea to a set of targets, creating the so called *scan path* [140], as shown in Figure 2.6.

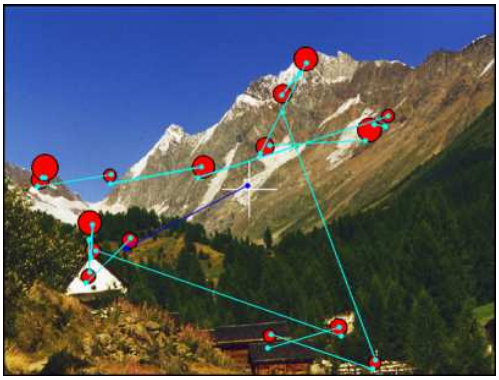


Figure 2.6: Example of a human scan path.

The question that arises here is how these saccades are generated. As early as 1967, Yarbus ascertained a relationship between saccadic eye movements and visual attention [165]. More recent works argued that visual attention anticipates an eye movement at the same scene location [135, 60, 50, 82]. Furthermore, it has been stated in [96] that attention can be shifted about four times before the next eye movement takes place. This behavior allows the attention mechanism to examine several targets and retain the most important one, to which the fovea is then shifted. Based on these works, the link between the shift of attention and eye movement has been formulated in a model, the Eye Movement and Movement of Attention (EMMA) [133], which considers visual attention as a predicting module for eye movements.

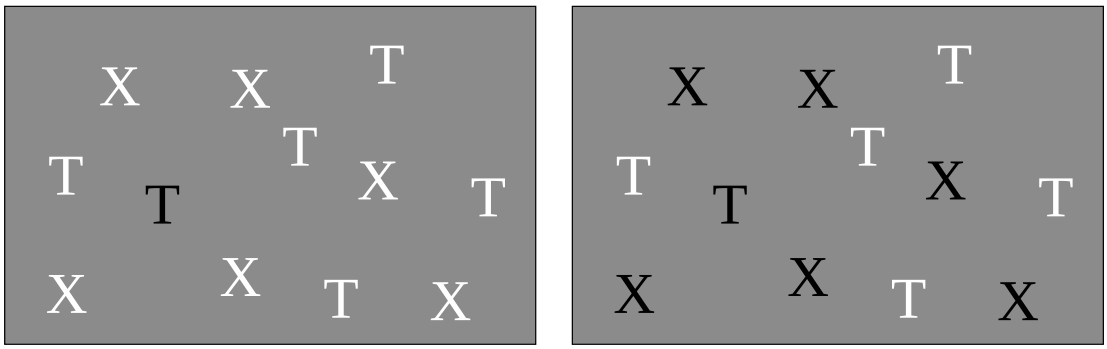
Note that the kind of attention deployment that causes eye movement is called *overt* attention (as opposed to covert attention which is not followed by an eye movement).

To summarize this section, a large part of the eye saccades is generated and controlled by the visual attention mechanism in order to foveate salient or informative locations of the observed scene. The next relevant question would be how the human visual attention mechanism selects the locations to be attended next.

2.2.4 Psychophysical Models of Visual Attention

This section deals with the question asked at the end of the previous section, namely what makes an image location more salient or more attractive than other locations. Answering this question means building a model of the human visual attention. Several psychophysical models of visual attention have been reported in the literature like the spotlight model reported in [102], the texton model presented by Julesz [76] and the zoom-lenz model [46]. A recent and excellent review of the most important psychophysical models of visual attention is given in [63].

Most of the reported models agree that the attention mechanism consists of two functionally independent stages: An early *preattentive* stage that operates in parallel across the entire visual field and a later *attentive* stage that deals with few scene parts sequentially.



(a) "find the black letter" (b) "find the black T"

Figure 2.7: Examples of the pop-out and conjunction problems in visual search.

In this perspective, Treisman proposed the *feature integration theory* that gave rise to one of the most popular models of visual attention [148]. The model was developed out of some experiments on human subjects about visual search, that is the task to find a known object (target) among many others (distractors). Indeed, the time needed by the human subjects to solve a visual search task (reaction time) is supposed to give insights to the visual attention behavior. These experiments revealed two different behaviors with respect to reaction time. First, if the target differs from all distractors (the other items) for at least one feature then, the reaction time of subjects is constant regardless of the number of distractors (see Figure 2.7 (a)). This kind of targets are said to *pop-out* of the visual field.

Second, if the target is distinguishable by a combination of features (Figure 2.7 (b)), the reaction time of subjects increases linearly with the number of distractors. That means that the subjects sequentially scan all items until the

target is found. During this scan the considered features are combined together in order to discriminate the target.

Further to these experiments, Treisman proposed a model of attention that consists of a preattentive and an attentive step. The preattentive step, which is inspired from the constant time search tasks, extracts a set of feature maps (color, orientation, ...) in a parallel way over the entire visual field. Each map retains the ability to detect activity in it, which permits the rapid decision whether the wanted stimuli is present.

The attentive step of the model which is inspired from the linear time search tasks acts serially on single scene items or locations [149]. For each item, the different features are analyzed with respect to the conjunction criteria and are mapped into a master map of locations [147]. This procedure is repeated until the target is identified.

The feature integration theory of Treisman served as a basis for the conception of numerous computational models of visual attention. The following section reports some of them.

2.3 Visual Attention in Machines

Due to the combinatorial aspect of numerous computer vision problems, the selection of a reduced set of the available sensory information for further processing is of fundamental importance to master the complexity issue in computer vision [150, 151]. The progress made in understanding the human visual attention mechanism has stimulated the computational modeling and implementation of this mechanism.

2.3.1 Bio-inspired Computational Models of Attention

As mentioned in the previous section numerous computational models of visual attention have been inspired by the work of Treisman and colleagues [148, 149]. The majority of these models are composed of the two main steps introduced above; a parallel preattentive step and a sequential attentive one.

The Saliency-based Model of Koch and Ullman

Koch and Ullman presented in [79] one of the most popular computational models of visual attention. The biologically plausible model is purely data-driven (bottom-up), which means that only the image data are considered to stimulate attention shifts. The model relies on three main principles, as illustrated in Figure 2.8:

- The saliency of locations over the entire visual field must be represented in a unique scalar map, called the saliency map (which corresponds to the

master location map of the feature integration theory).

- The saliency of scene locations is strongly influenced by the surrounding context.

- Winner-Take-All (WTA) and the inhibition of return are suitable mechanisms to allow the attention deployment over the visual field.

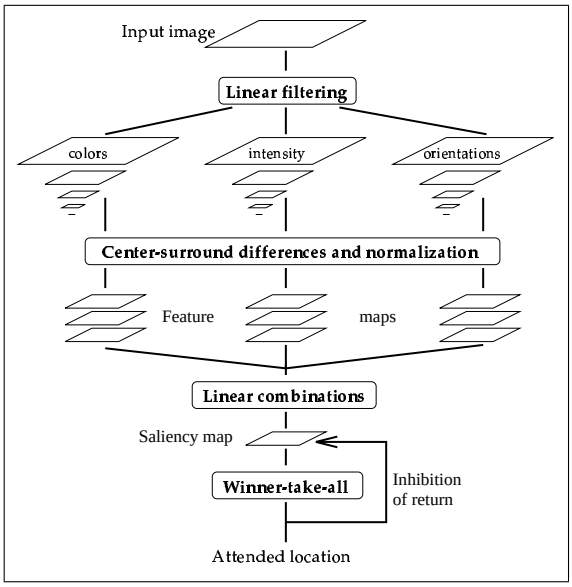


Figure 2.8: The saliency-based model of visual attention as suggested by Koch and Ullman (from [74]).

Indeed, the model starts with extracting, in a parallel manner, a set of scene features like color, orientation, motion etc. Always in a parallel way, a conspicuity map is computed for each considered feature using a lateral inhibition mechanism. Thus, each conspicuity map highlights the parts of the scene that strongly differ, according to the corresponding feature, from their surroundings. Koch and Ullman suggested that multiscale center-surround filters are suitable means to implement the conspicuity transformation.

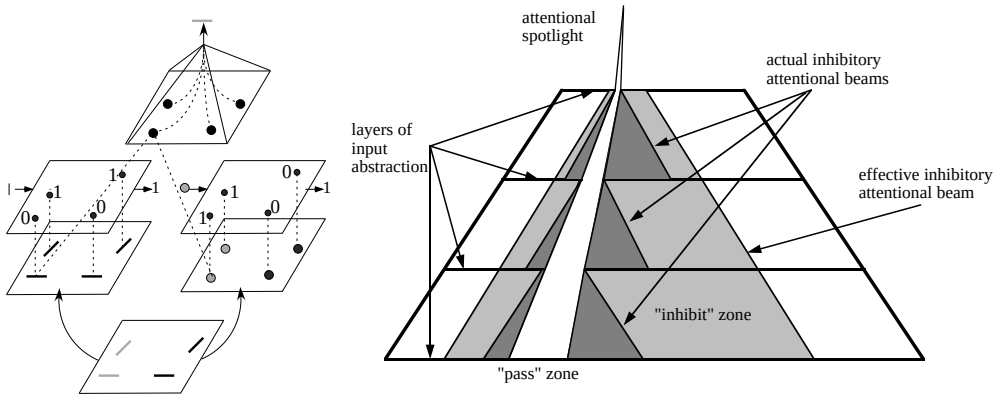
The different conspicuity maps are then merged into a single map of attention called the saliency map which topographically encodes for location saliency over the entire scene and with respect to all considered features.

Given a saliency map, a further step of the model consists in finding the most salient scene locations. A maximum network which is generally implemented using a Winner-Take-All (WTA) network allows the selection of the most salient locations often referred to as Focus of Attention (FOA).

Related Models

Several computational models of visual attention have been then derived from the model developed by Koch and Ullman.

The first attempt to implement the model proposed by Koch and Ullman was presented by Chapman in [28]. Thereby, Chapman intended to computationally reproduce the visual search results reported by Treisman in [149]. Like the saliency-based model, Chapman's starts with computing a set of feature maps. Second, each feature map is transformed into a binary activation map which marks the relevant locations. Unlike the saliency-based model, the conspicuity transformation which transforms a feature map into an activation map is based on thresholds rather than lateral inhibition. At this stage of his model, Chapman proposed two working modes of his system; one simulates parallel search (for pop-out stimuli), and the other mode simulates the serial search (for conjunction stimuli). The parallel search is implemented as a global OR which combines all stimuli at the activation map related to the feature that uniquely discriminates the target. For the conjunction stimuli a serial examination of items is necessary. However, Chapman proposed a suitable hierarchical representation of the stimuli and the corresponding features in order to reduce the number of items to be serially examined for conjunction criteria (see Figure 2.9(a)).



(a) Champman's model (b) Selective tuning model

Figure 2.9: Models of attention inspired from the saliency-based one. (a) Chapman's model (from [28]). (b) The selective tuning model of attention (from [38]).

Visual attention via selective tuning reported in [37, 38, 152] has been also largely inspired from the ideas proposed by Koch and Ullman. However, the model does not cover the computation of a saliency map. Supposing that such a map exists, the authors have concentrated their effort to develop a consistent Winner-Take-All mechanism. They have introduced the concept of a processing hierarchy which can be seen as a pyramidal representation of the saliency map.

At each layer of the hierarchy, the elements are computed as a weighted sum of certain neighbors from the underlying layer. At the top level of the pyramid a "beam", with a certain radius, is projected around a winner location (e.g. the location with the highest activity) and expands as it traverses the hierarchy. A WTA mechanism is activated at each level in order to "route" the beam into the next level. The global winner is then determined at the lowest level of the hierarchy (see Figure 2.9(b)).

This model is seen by other authors as a consistent implementation of the WTA mechanism rather than a complete visual attention system, since the computation of the saliency map has not been considered [103, 69].

Milanese is among the first who investigated the implementation of the full model of visual attention suggested by Koch and Ullman. Indeed, in his elaborate and detailed PhD thesis [96], he deeply discussed the implementation of the different steps of the model. A set of feature maps related to color, intensity and orientation are first extracted from a color input image. Second, each feature map is transformed into a conspicuity map that discriminates outstanding regions according to a specific feature, using the center-surround mechanism. Milanese used Difference of Oriented Gaussian filters ($\mathcal{D}o\mathcal{O}r\mathcal{G}$) to implement the multiscale conspicuity transformation. Finally, the conspicuity maps are integrated into the final saliency map, after undergoing a non-linear relaxation process [97]. The model does not, however, implement the WTA mechanism to, successively, select the most salient locations of the scene.

The major disadvantage of the model proposed by Milanese lies in its computation complexity which is not due to basic conception of the model, but to the way this concept was implemented. Indeed, Milanese did not take advantage of the multiresolution representation of feature maps to implement the multiscale conspicuity transformation. Instead, he applied a filter bank with variable sizes on fixed size images. In a study comparing the implementation of Milanese with the one of Itti, a colleague of Koch at Clatech, who used the multiresolution concept to implement the same model, we concluded that the latter implementation is about 1000 times faster than the former one [113].

A detailed description of the multiresolution implementation of the saliency-based model of attention reported by Itti in [74] will be given in Section 2.4.

Given its completeness and efficiency, this multiresolution implementation of the saliency-based model of attention is adopted in this thesis.

Other Models

The previous section reported some computational models which are closely related to the saliency-based model of Koch and Ullman. There exist, however, other works that have dealt with the computational modeling of visual attention. For instance, VISIT is a connectionist model of attention which involves bottom-up as well as top-down mechanisms to select interesting objects in a scene [2].

The Guided Search model [164, 25] is another example that integrates image-based stimuli and task-dependent knowledge into an overall *activation map* which corresponds to the saliency map. In a recent work [89], a very similar model of attention with a top-down component has been reported.

Olshausen *et al.* have presented a computational model that simulates the *shifter circuits theory* [109], that is the segregation of objects of interest and the routing of the corresponding visual information to higher stage of the visual cortex.

In a recent work Privitera and Stark [126] have proposed a model of visual attention, whose internal parameters (used features, feature weights,...) can be adapted to the type of the analyzed image. The purpose of the work was to reproduce the human scan path by a computational model.

A more complete review on the existing computational models of visual attention is presented in [96] and [63].

2.3.2 Visual Attention and Computer Vision Applications

The development of computational models of visual attention, their realization using standard computer vision techniques and the need for speeding up computer vision tasks encouraged researchers to integrate this mechanism in their computer vision applications. The emergence of active and purposive computer vision [4, 3] has increased the relevance of visual attention in computer vision.

Applications like character recognition and robot navigation, among several others, largely benefitted from the consideration of the visual attention paradigm as a component of computer vision solutions.

Object Recognition

Model-based object recognition which consists in finding correspondences between image and model features is basically a combinatorial problem. The visual attention paradigm can reduce the complexity of the recognition task by reducing the amount of image data to be processed by this task.

One of the earliest works that has integrated a visual attention module in an object recognition system has been presented in [100]. The Multiple Object Recognition and attentional SELection (MORSEL) model is a connectionist system conceived for the recognition of simple 2D objects in an image. Most of the results presented in [100, 99] have been carried out to recognize letters and words.

Another work that used visual attention in the field of character recognition was reported in [5] and extended in [132]. The complete system is composed of three levels, the *attentive level*, the *intermediate level* and the *associative level*. The attentive module is similar, in its structure, to the model of Koch and Ullman. A saliency map is computed using simple features (the author implemented only line orientation). A WTA network detects then the most informative parts of the

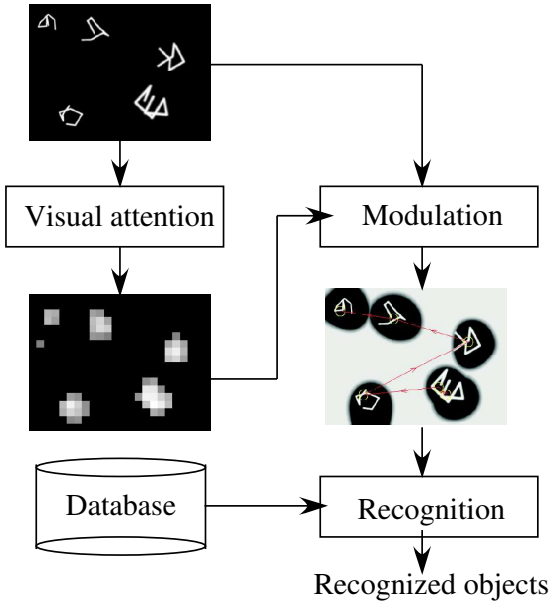


Figure 2.10: Attentive object recognition (after [159]).

image, which are passed to the intermediate module that analyzes the content of each selected part. Finally, the associative module combines the information about all selected locations in order to recognize the objects present in the image. Originally, this system was used for character recognition, but was also extended to recognize faces.

Recently, the saliency-based model of visual attention has been integrated into an object recognition system [159]. The main goal of the attention module was to provide the recognition system with a first order approximation about the location and extent of most salient image regions. Thus, instead of trying to interpret the entire scene, the recognition module focuses on the scene parts previously provided by the visual attention module. Note that the object recognition module, HMAX [129], used in that work has been conceived to mediate object recognition in cortex. indeed, it has been applied on bar-like stimuli in order to account for the experimental data of Logothetis *et al.* on object representation in monkeys [88]. A schematic description of the complete system is illustrated in Figure 2.10.

Active Vision Systems

Aloimonos defined an active observer as the one who is capable of engaging in some kind of activity whose purpose is to control the geometric parameters of its sensory apparatus [3]. Thus, an active computer vision system should dispose of a module that controls the pan and tilt parameters of its cameras, i.e. a *gaze*

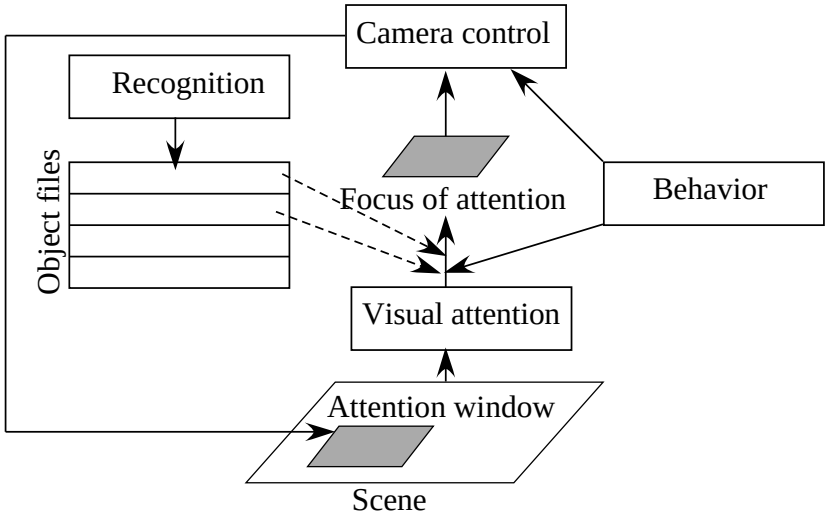


Figure 2.11: Overview of the NAVIS system (from [10]).

control module. The visual attention paradigm has been recognized as a powerful tool to guide these camera movements.

One of the earliest attempts to use visual attention as a gaze control mechanism in an active vision system has been achieved by Clark and Ferrier [31]. The system was developed on a binocular camera head whose pan, tilt and vergence parameters are dynamically controlled in real time. It performs pursuit task and saccade generation task, independently. Visual attention was responsible for the saccade generation task and uses learned templates of the environment to extract the feature maps. The integration of the different features into an attention map is task dependent, since learned weights are assigned to each feature type. The activities of the attention map determines the location to which the camera head should be oriented next.

A similar active system has been developed at the KTH Stockholm by the group of Eklundh [23]. The gaze control mechanism involves a visual attention model which is based on two main features, namely depth and motion [92, 93]. Like the system of Clark and Ferrier, the authors implemented two functioning modes; a pursuit mode that tracks already selected objects and a saccade mode that shifts attention to another moving object that newly entered the visual field.

A more recent and especially more complete work on the integration of visual attention into an active vision system has been carried out in the IMA Lab (university of Hamburg) [17, 10]. The complete system NAVIS (Neural Active VISion) can be roughly divided into two components: an attention component and a camera control one. The attention module is built around features like color, edge symmetry and region eccentricity. A combination of these features and some top-down information about the environment give rise to the master

attention map. The salient scene locations are then selected and transferred to the camera control module in order to perform a gaze shift to that locations. Figure 2.11 gives an overview of the active vision system NAVIS.

Other works have used visual attention in other active vision systems, especially to help mobile robot to navigate in known or unknown environments, like in [43, 146, 163], to cite only few of them.

2.4 The Saliency-based Model of Visual Attention

As already mentioned, the saliency-based model of visual attention has been presented by Koch and Ullman in [79]. It is based on four major principles: visual attention acts on a multi-featured input; saliency of locations is influenced by the surrounding context; the saliency of locations is represented on a scalar map: the saliency map; and the Winner-Take-all and inhibition of return are suitable mechanisms to allow attention shift.

Itti *el al.* have presented a complete implementation of the saliency-based model in [73]. In the following, the implementation details of the four main steps of the model are presented (see Figure 2.12).

2.4.1 Feature Maps

First, a number of features $(1..j..n)$ are extracted from the scene by computing the so called feature maps F_j. Such a map represents the image of the scene, based on a well-defined feature, which leads to a multi-featured representation of the scene. In his implementation, Itti considered seven different features which are computed from an RGB color image and which belong to three main cues, namely intensity, color, and orientation.

- Intensity feature

$$F_1 = I = 0.3 \cdot R + 0.59 \cdot G + 0.11 \cdot B \tag{2.1}$$

- Two chromatic features based on the two color opponency filters R^+G^- and B^+Y^- where the yellow signal is defined as $Y = \frac{R+G}{2}$. Such chromatic opponency exists in human visual cortex [45].

$$\begin{aligned} F_2 &= \frac{R-G}{I} \\ F_3 &= \frac{B-Y}{I} \end{aligned} \tag{2.2}$$

  The normalization of the chromatic features by I decouples hue from intensity.

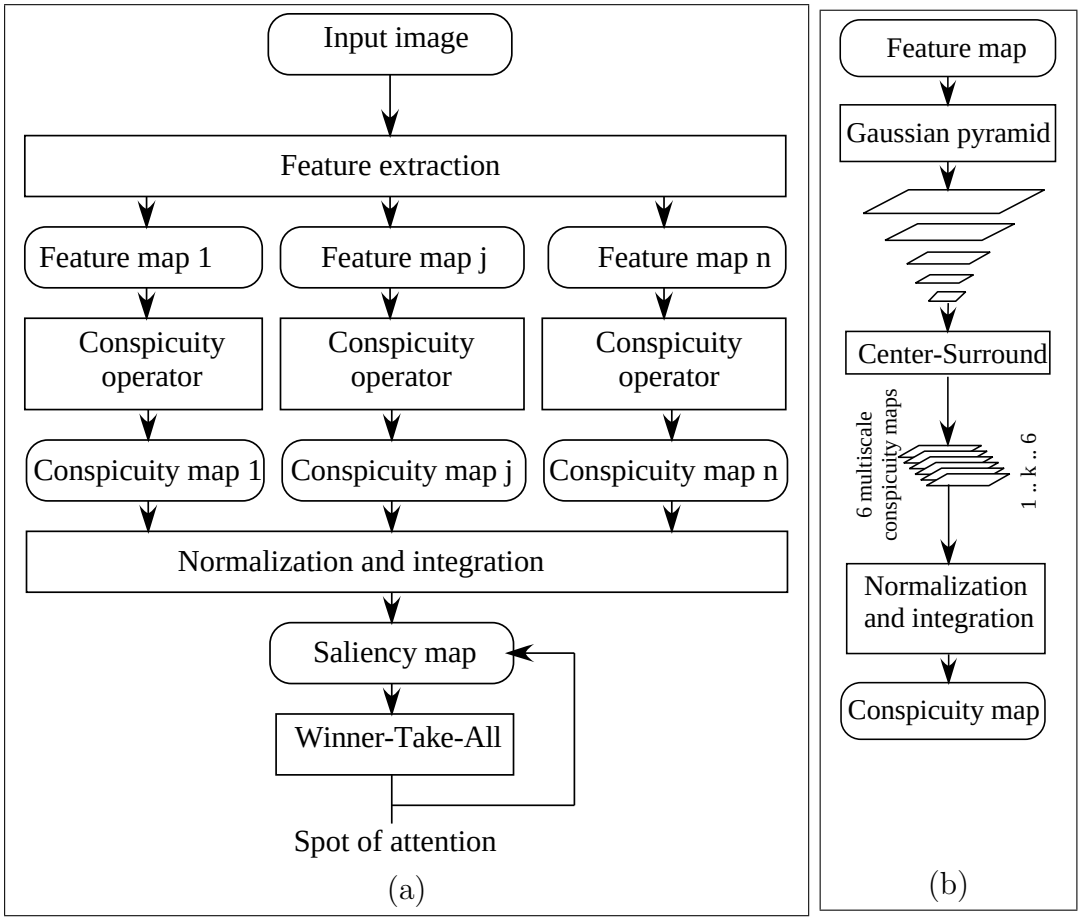


Figure 2.12: Saliency-based model of visual attention. (a) represents the four main steps of the visual attention model. Feature extraction, conspicuity computation (for each feature), saliency map computation by integrating all conspicuity maps and finally the detection of spots of attention by means of a winner-take-all network. (b) illustrates, with more details, the conspicuity operator, which computes six intermediate multiscale conspicuity maps. Then, it normalizes and integrates them into the feature-related conspicuity map.

- Four local orientation features $F_{4..7}$ according to the angles $\theta \in \{0^o, 45^o, 90^o, 135^o\}$. Gabor filters, which represent a suitable mathematical model of the receptive field impulse response of orientation-selective neurons in primary visual cortex [85], are used to compute the orientation features. In this implementation of the model, it is possible to use an arbitrary number of orientations. However, it has been noticed in [73] that using more than four orientations does not improve the performance of the model drastically.

2.4.2 Conspicuity Maps

In a second step, each feature map is transformed in its conspicuity map which highlights the parts of the scene that strongly differ, according to a specific feature, from their surroundings. In biologically plausible models, this is usually achieved by using a *center-surround*-mechanism. Practically, this mechanism can be implemented with a *difference-of-Gaussians*-filter, $\mathcal{D}o\mathcal{G}$, which can be applied on feature maps to extract local activities for each feature type.

A visual attention task has to detect conspicuous regions, regardless of their sizes. Thus, a multiscale conspicuity operator is required. It has been shown in [96], that applying variable size center-surround filters on fixed size images, has a high computational cost. An alternative method has been presented in [74]. This method is based on a multi-resolution representation of images. For a feature j, a gaussian pyramid $\mathcal{P}_j$ is created by progressively lowpass filtering and subsampling by factor 2 the feature map F_j, using a gaussian filter G:

$$\begin{aligned} \mathcal{P}_j(0) &= F_j \\ \mathcal{P}_j(i) &= \left(\downarrow 2\right)(\mathcal{P}_j(i-1) * G) \end{aligned} \tag{2.3}$$

where $(*)$ refers to the spatial convolution operator and $\left(\downarrow 2\right)$ refers to the dowsampling operation. Center-Surround is then implemented as the difference between fine (c for center) and coarse scales (s for surround). Indeed, for a feature j $(1..j..n)$, a set of intermediate multiscale conspicuity maps $\mathcal{M}_{j,k}$ $(1..k..K)$ are computed according to Equation 2.4, giving rise to $(n \cdot K)$ maps for n considered features.

$$\mathcal{M}_{j,k} = |\mathcal{P}_j(c_k) \ominus \mathcal{P}_j(s_k)| \tag{2.4}$$

where $\ominus$ is a cross-scale difference operator that first interpolates the coarser scale to the finer one and then carries out a point-by-point substraction.

The absolute value of the difference between the center and the surround allows the simultaneous computing of both sensitivities, dark center on bright surround and bright center on dark surround (red/green and green/red or blue/yellow and yellow/blue for color).

It is noteworthy that these intermediate multiscale conspicuity maps are sensitive to different spatial frequencies. Fine maps detect high frequencies and thus small image regions, whereas coarse maps detect low frequencies and thus large regions.

For the orientation features, oriented Gabor pyramids [59] are used instead of the gaussian ones (for more details about Gabor pyramids, see A.2).

Next, for each feature j, the multiscale maps $\mathcal{M}_{j,k}$ are combined, in a competitive way into a unique feature-related conspicuity map C_j in accordance with Equation 2.5.

$$C_j = \sum_{k=1}^{K} \mathcal{N}(\mathcal{M}_{j,k}) \tag{2.5}$$

where $\mathcal{N}(.)$ is a normalization operator which simulates the competition between the different scales. A detailed description of different normalization strategies is given in Section 2.4.5.

Note that the summation of the multiscale maps is achieved at the coarsest resolution. Maps of finer resolutions are lowpass filtered and downsampled to the required resolution.

Finally, the n conspicuity maps C_j, are combined into cue-related conspicuity maps $\hat{C}_{cue}$ according to Equation 2.6. Each $\hat{C}_{cue}$ regroups the different conspicuity maps $C_{j_{cue}}$ which belong to the same cue (intensity, color, orientation ...).

$$\hat{C}_{cue} = \sum_{j_{cue}} \mathcal{N}(C_{j_{cue}}) \tag{2.6}$$

In fact, it has been pointed out in [149] that features of the same cue directly compete for saliency (intra-cue competition), whereas features of different cues can only compete through cue integration (cross-cue competition).

2.4.3 Saliency Map

In the third step of the attention model, the cue-related conspicuity maps $\hat{C}_{cue}$ are integrated together, in a competitive manner, into a *saliency map* $\mathcal{S}$ in accordance with Equation 2.7.

$$\mathcal{S} = \sum_{cue=1}^{l} \mathcal{N}(\hat{C}_{cue}) \tag{2.7}$$

where l is the number of the considered cues. As mentioned above, the normalization operator $\mathcal{N}(.)$ is described in Section 2.4.5.

In his implementation [74], Itti used seven features ($n = 7$) which belong to three different cues ($l = 3$). Regarding the conspicuity operator, he computed six ($K = 6$) multiscale conspicuity map $\mathcal{M}_{j,k}$ (a detailed description of the six maps is given in A.1). Thus, for the seven features he computed 42 multiscale maps.

2.4.4 Selection of Salient Locations

At any given time, the maximum of the saliency map defines the most salient location, which represents the actual spot of attention. A "winner-take-all" (WTA) mechanism [79] is used to detect, successively, the significant regions. Given a saliency map, the WTA mechanism starts with selecting the location with the maximum value of the map. This selected region is considered as the most salient part of the image (winner). The spot of attention is then shifted to this location. Local inhibition is then activated in the saliency map, in an area around the actual spot. This yields dynamical shifts of the spot of attention by allowing the next most salient location to subsequently become the winner. Besides, the inhibition mechanism prevents the spot of attention from returning to previously attended locations. The number of detected locations can be either set by the user or determined automatically through the activities of the saliency map.

2.4.5 Normalization Strategies for Map Combination

The saliency-based model of visual attention performs two kinds of map combination. On one hand, the cross-scale combination of the multiscale conspicuity maps $\mathcal{M}_{j,k}$ in order to compute a unique conspicuity map C_j for each scene feature. On the other hand, the cross-feature and the cross-cue combination that integrates the different feature-related conspicuity maps into the final saliency map. The combination of a set of maps should, in the context of visual attention, simulate a competition for saliency between different scales, but also between different scene features.

Two interesting bottom-up combination strategies have been presented in [71].

Contents-based Global Amplification Normalization ($\mathcal{N}_1(.)$)

Given conspicuity maps which should be integrated into a unique map, the normalization strategy $\mathcal{N}_1(.)$ consists in the following:

1. Scale all maps to the same dynamic range in order to eliminate across-modality amplitude difference due to dissimilar extraction mechanisms.

2. For each map, compute the global maximum M and the average $\overline{m}$ of all the other local maxima. A local maximum of a map is defined as a location whose value is larger than those of its direct neighbors.

3. Globally multiply the map by a weight $w_{\mathcal{M}} = (M - \overline{m})^2$. Thus, $\mathcal{N}_1()$. normalizes a conspicuity map $\mathcal{M}$ in accordance with Equation 2.8.

$$\mathcal{N}_1(\mathcal{M}) = w_{\mathcal{M}} \cdot \mathcal{M} \tag{2.8}$$

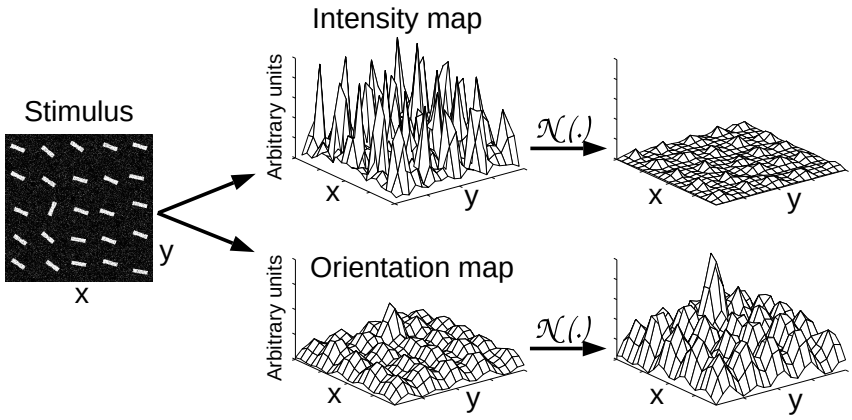


Figure 2.13: Contents-based global amplification normalization. The intensity map contains comparable responses which leads to a small w. For this reason the intensity map is strongly suppressed. Due to the presence of an outstanding location in the orientation map, the corresponding w is large, which explains the global amplification of that map (from [74]).

In fact, w measures how the most active locations differ from the average of local maxima of a conspicuity map. Thus, this normalization operator promotes conspicuity maps in which a small number of strong peaks of activity is present. Maps that contain numerous comparable peak responses are demoted. This effect is clearly illustrated in Figure 2.13. It is obvious that this competitive mechanism is purely data-driven and does not require any a priori knowledge about the analyzed scene.

Iterative Non-Linear Normalization ($\mathcal{N}_2(.)$)

The non-linear normalization strategy $\mathcal{N}_2(.)$ is composed of the following steps. First, all maps are normalized to the same dynamic range in order to remove modality-dependent amplitude differences. Second, each map is iteratively convolved by a large 2D $\mathcal{D}o\mathcal{G}$ filter. The negative results are clamped to zero after each iteration, which represents the non-linearity of this normalization method. At each iteration of the normalization process, a given map $\mathcal{M}$ is transformed in accordance with Equation 2.9.

$$\mathcal{M} \leftarrow |\mathcal{M} * \mathcal{D}o\mathcal{G}|_{\geq 0} \tag{2.9}$$

where $(*)$ is the convolution operator and $|.|_{\geq 0}$ discards negative values.

The normalization strategy $\mathcal{N}_2(.)$ relies on simulating local competition between neighboring conspicuous locations. Spatially grouped locations which have similar conspicuities are suppressed, whereas spatially isolated conspicuous lo-

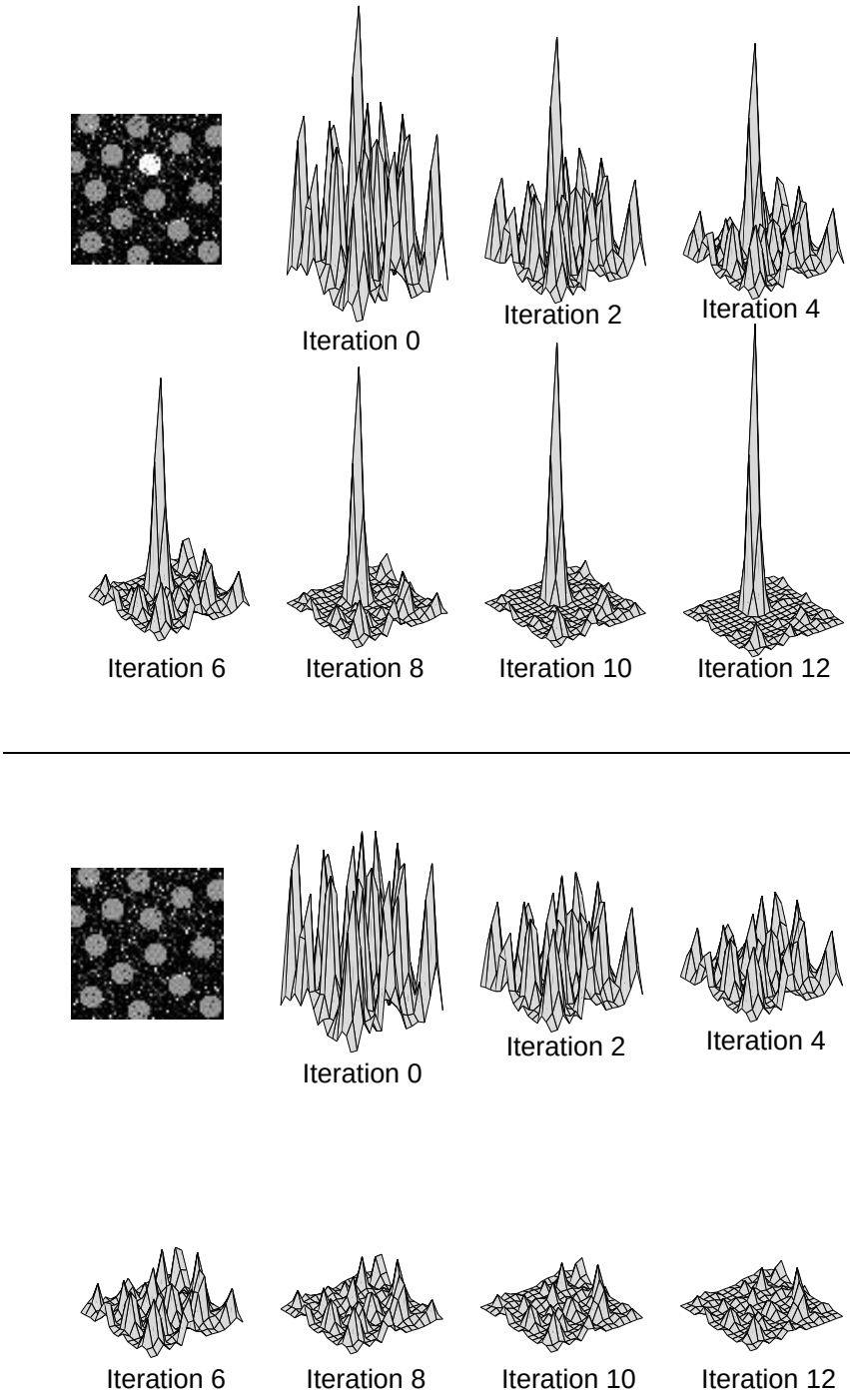


Figure 2.14: Iterative non-linear normalization. The upper example illustrates how the non-linear normalization progressively promotes the major peak while suppressing less conspicuous locations. The bottom example shows the suppressing of the entire map in the absence of outstanding peaks (from [72]).

cations are promoted. The behavior of the iterative non-linear normalization method is illustrated in Figure 2.14.

In addition to the fact that it is inspired from the human vision [73], this normalization strategy, thanks to its non-linearity, has the advantage to suppress the noise of a conspicuity map, while promoting the major peaks in the same map.

It has, however, a practical drawback regarding its complexity. Indeed, iteratively applying a large filter on an image is time consuming, unless the size of the filter does not intervene in the complexity of the convolution operator (e.g. recursive filters). The choice of the number of iterations represents also a drawback of this normalization method.

We will see in the coming chapters that both normalization strategies are useful, depending on the available computation resources and on the applications.

2.5 Chapter Summary

This chapter can be summarized in the following points:

- Visual attention is an essential mechanism for the human visual system, since it permits the selection of the visually-relevant parts of the scene, on which foveated and high level vision is performed. Also, the visual attention paradigm partially explains the high performance of our vision system.

- Visual attention is a fundamental tool for computer vision, if we want to overcome the complexity of vision tasks. Numerous bio-inspired computational models of visual attention have been presented in literature. The saliency-based model of attention represents the most admitted model that simulates human visual attention.

Chapter 3

Extensions of the Basic Model of Visual Attention

3.1 Chapter Introduction

As mentioned in Chapter 2, most of the studies that proposed an implementation of the saliency-based model of visual attention considered static 2D features like color and orientation to compute the saliency map. Little attention has been devoted so far to 3D and dynamic scene features as source of visual attention.

This chapter presents our effort to integrate 3D and dynamic scene features into the saliency-based model of visual attention.

3.1.1 Chapter Outline

The current chapter is divided into two main parts. The first part (Section 3.2) deals with the integration of 3D-related features into the model of visual attention. Therefore, the acquisition of scene depth using the depth from stereo principle is first described. Then, the relevance of some depth-related features to the attention mechanism is investigated. Finally, the combination of 2D- and 3D-related features into the final saliency map is presented, and some illustrative examples are given.

Covering the second part of this chapter, Section 3.3 reports a model of dynamic visual attention that combines static and dynamic features to detect salient objects in dynamic scenes. The section starts with reporting some motion estimation techniques. Then, the computation of the required dynamic conspicuity map, using a hierarchical gradient-based motion estimation method, is presented. After a description of the integration of static and dynamic conspicuity maps into the final saliency map, experiments that validate the proposed model are finally reported.

3.2 Model of Visual Attention for 3D Vision

It is generally admitted that depth plays a significant role in early stages of the human visual system [95], which reinforces the hypothesis that this cue strongly contributes to visual attention. In computer vision, it is agreed that depth is a highly relevant feature in scene analysis and the early selection of scene parts based on this feature can strongly influence the performance of higher level vision tasks [35]. Furthermore, the present availability of 3D range cameras makes it possible to implement and to validate visual attention models that consider also 3D features, which was not possible few years ago.

3.2.1 Depth from Stereo

Although the computation of depth itself is out of the scope of this thesis, I think that a short introduction to the existing depth measurement techniques could be helpful for the understanding of this part of the report.

Stereoscopic vision refers to the derivation of depth information from a pair of images. Indeed, the slightly different perspectives from which a pair of horizontally aligned cameras observe a scene leads to different images with relative displacements of objects (disparities) in the two monocular views of the scene. The size of the disparities of an object is a measure of its relative depth, from which the absolute depth-information can be obtained if the geometry of the imaging system is known.

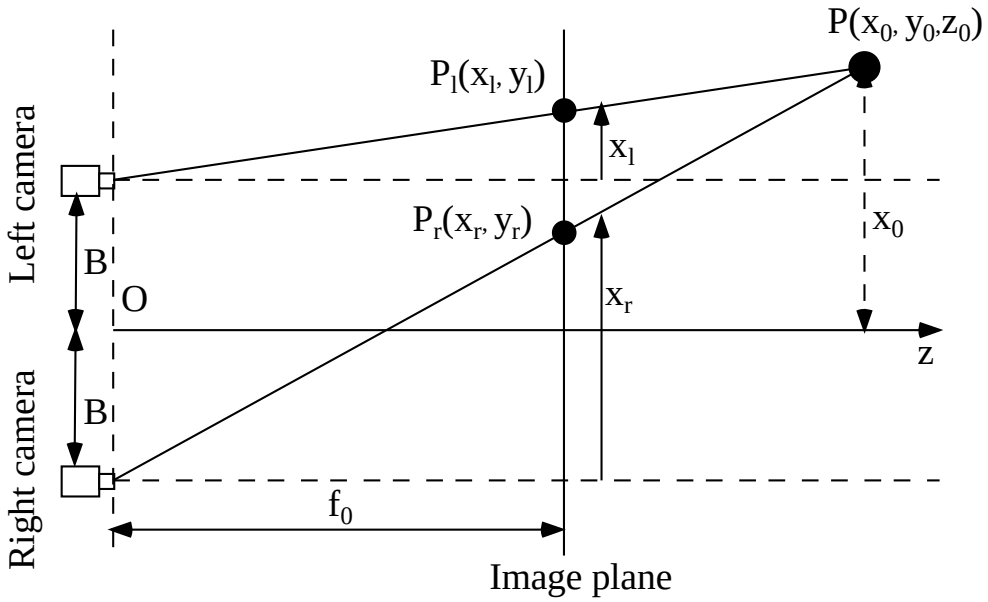


Figure 3.1: Principle of stereoscopic vision.

The principle of stereoscopic vision is depicted in Figure 3.1. Let $2B$ (baseline) be the distance separating two cameras having the same focal length f_0 and parallel optical axes. A physical point $P(x_0, y_0, z_0)$ is imaged to $P_l(x_l, y_l)$ on the image plane of the left camera and to $P_r(x_r, y_r)$ on the image plane of the right one. By means of geometry, the coordinates x_l and x_r can be computed according to Equation 3.1 [48].

$$\begin{aligned} x_l &= \frac{f_0(x_0 - B)}{z_0} \\ x_r &= \frac{f_0(x_0 + B)}{z_0} \end{aligned} \tag{3.1}$$

The depth z_0 of the point P can be then derived in accordance with Equation 3.2.

$$z_0 = \frac{2f_0B}{d} \tag{3.2}$$

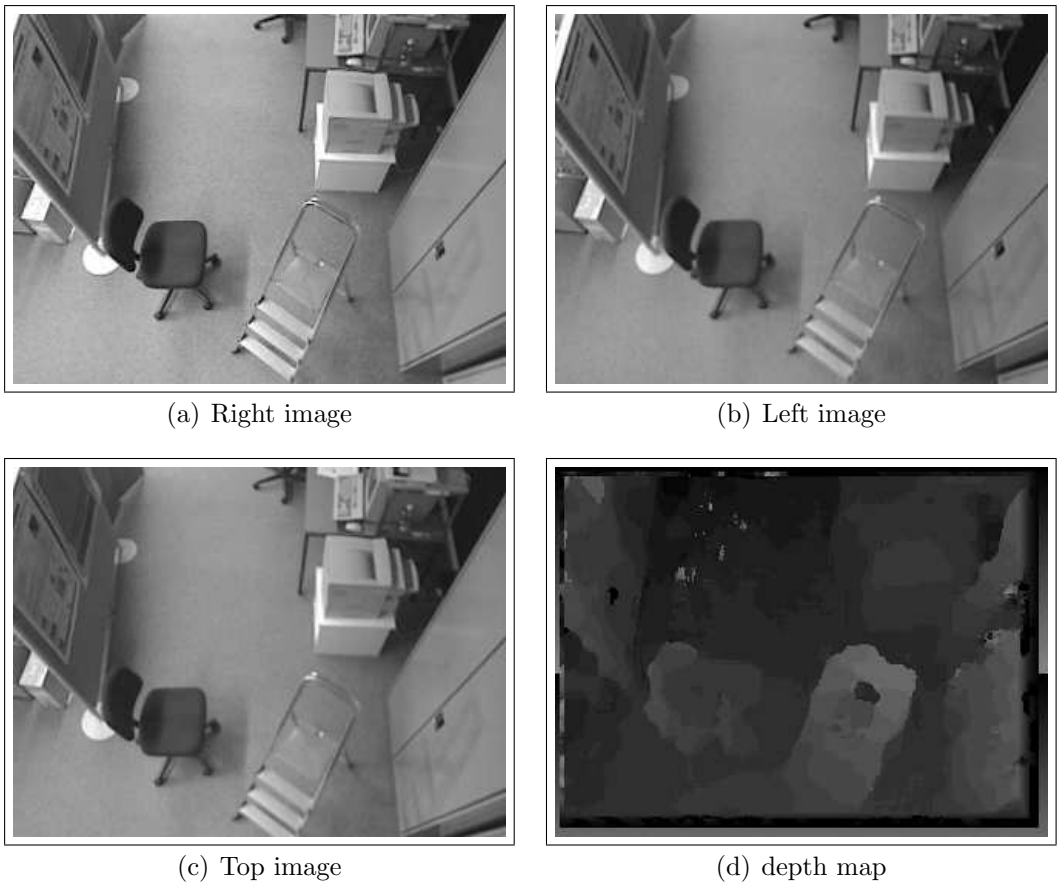
where the disparity d is the difference between x_r and x_l ($d = (x_r - x_l)$).

A major challenge in stereoscopic vision is the solving of the correspondence problem which consists in finding, for each pixel of the left image the corresponding pixel in the right image. Numerous works have dealt with this problem and proposed novel solutions [9, 77, 49, 141], which stimulated the development of 3D stereo devices.

Recently, the company Point Grey Research [128] developed a 3D vision system named Triclops, which consists of a three-camera module and a software system that performs depth measurements. A picture of Triclops is given in Figure 3.2.

Figure 3.2: Triclops.

Triclops computes scene depth from three calibrated cameras using an algorithm similar to the multi-baseline stereo [108]. Thus, two camera pairs are available, a classical horizontal pair (left, right) and a vertical pair (right, top). The trinocular cameras can achieve better results than a typical two camera stereo system because the vertical pair of cameras can resolve situations that are ambiguous to the horizontal pair. Figure 3.3 illustrates a depth map computed from three images (left, right, top) acquired by Triclops.



(a) Right image (b) Left image

(c) Top image (d) depth map

Figure 3.3: Stereo image: Triclops.

In addition to the depth map, Triclops provides also the associated color image of the scene.

3.2.2 Conspicuity from Depth-Related Features

Given 3D information represented by the depth map, the goal of this section is to extract depth-related features which are relevant to the attention mechanism. We investigated three different features [114], namely depth itself, mean curvature,

and depth gradient.

Depth Feature

As mentioned above, depth information seems to play a significant role in early human vision and, thus, in guiding attention. In computer vision, the distance between the sensor and the scene objects is a precious information for numerous tasks, such as obstacle avoidance. Thus, it can be concluded that this feature is of high relevance for visual attention. This feature is directly available from the sensor data and thus no additional operations are needed to compute it.

Mean Curvature

The mean curvature is an intrinsic surface feature that provides useful information about the geometry of the scene objects. Let $z(\mathbf{x}) = z(x, y)$ be the two-dimensional depth function, the mean curvature $H(\mathbf{x}) = H(x, y)$ can be formally defined at each location $\mathbf{x} = (x, y)$ in accordance with Equation 3.3 [15].

$$H(\mathbf{x}) = \frac{1}{2} \frac{(1 + z_x^2(\mathbf{x}))z_{yy}(\mathbf{x}) + (1 + z_y^2(\mathbf{x}))z_{xx}(\mathbf{x}) - 2z_x(\mathbf{x})z_y(\mathbf{x})z_{xy}(\mathbf{x})}{(1 + z_x^2(\mathbf{x}) + z_y^2(\mathbf{x}))^{\frac{3}{2}}} \tag{3.3}$$

where

$$\begin{aligned} z_x &= \frac{dz}{dx} \\ z_y &= \frac{dz}{dy} \\ z_{xy} &= \frac{d^2z}{dxdy} \end{aligned} \tag{3.4}$$

As a second order differential feature, mean curvature has, however, a remarkable disadvantage, i.e. its sensitivity to noise and non significance on depth discontinuities. The disadvantage related to noise sensitivity can be overcome through applying smoothing operators on the depth map. Depth discontinuities have to be detected in order to compute mean curvature only for continuous surfaces. Thus, integrating mean curvature into the computational model of visual attention requires some additional preprocessing operations.

Depth Gradient

This feature vector which is based on first order derivative of the depth function $z(\mathbf{x})$ can be an efficient means to detect important depth changes in the scene. In this work, we consider the magnitude $|\nabla z|$ of the gradient vector as defined in Equation 3.5.

$$|\nabla z(\mathbf{x})| = \sqrt{z_x^2(\mathbf{x}) + z_y^2(\mathbf{x})} \tag{3.5}$$

Figure 3.4 illustrates some observations about the significance of these various features. The first scene (scene 1) contains no depth discontinuities, but important curvature variation. Thus, mean curvature contributes strongly to the computation of the saliency map. The second experiment (scene 2) shows the sensitivity of mean curvature to depth discontinuities. It also shows that the image locations highlighted by the depth gradient are very similar to those locations highlighted by depth itself.

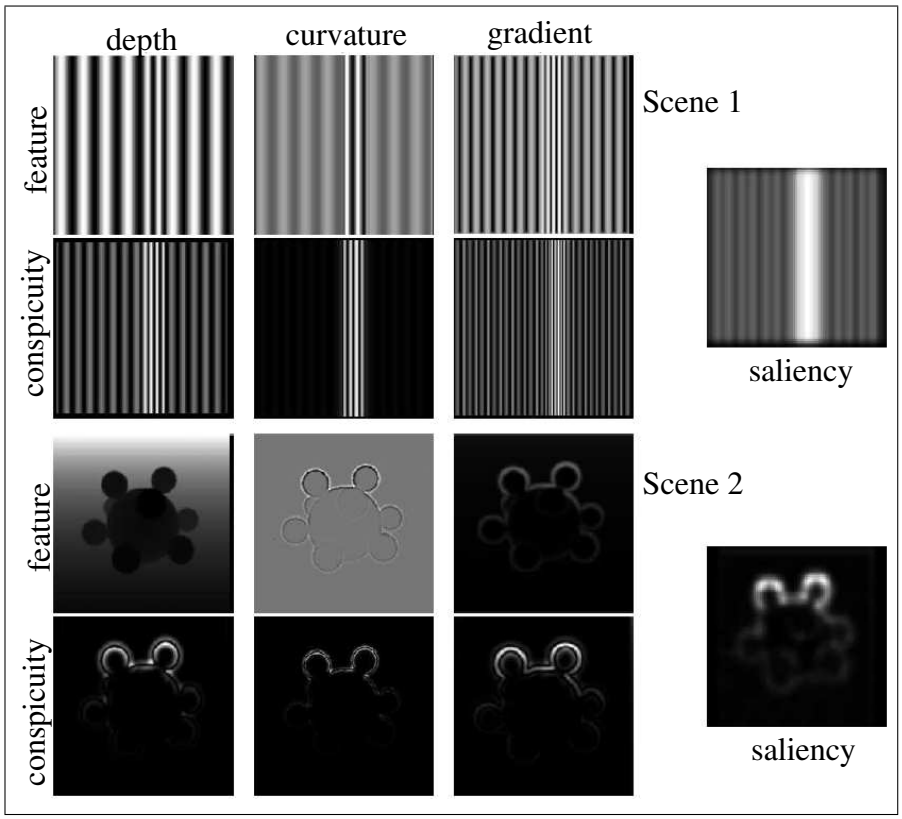


Figure 3.4: Conspicuity from depth-related features.

In fact, preliminary results tend to show a ranking of the features which is depth, mean curvature, depth gradient in an order of decreasing usefulness. A more deep experiments and analysis than those presented above are necessary in order to definitely assess the significance of mean curvature and depth gradient for the visual attention mechanism. However, the depth feature seems to be of high significance for visual attention as it will be demonstrated below.

3.2.3 Combining 2D- and 3D-Related Conspicuity Maps

The basic idea is to simply extend the saliency-based model of visual attention, described in Chapter 2.4, to the scene depth component. Given m suitable fea-

tures related to depth, the integration process can be achieved as follows. First, the feature maps are extracted from the depth data. The corresponding conspicuity map is then computed according to the same scheme described for 2D features in Section 2.4.2. Hence, in addition to the l cue-based conspicuity maps $\hat{C}_{cue}$ computed from the color image, an additional one $\hat{C}_{l+1}$, related to depth, is available. The integration module has to combine $l+1$ cue-based conspicuity maps in order to compute the saliency map. Equation 2.7, which has been used in classical models to compute the saliency map $\mathcal{S}$, can be adapted to Equation 3.6.

$$\mathcal{S} = \sum_{cue=1}^{l+1} \mathcal{N}(\hat{C}_{cue}) \tag{3.6}$$

3.2.4 Results and Discussion

This section presents some specific experiments carried out in order to validate the depth enhanced computational model of visual attention and to show the usefulness of depth information in a visual attention task. Two cues are considered in these experiments, *color* and *depth*. The color cue is composed of two features ($R-G$ and $B-Y$), whereas the depth cue is represented by one feature, namely depth itself. To achieve the cross-scale and the cross-feature combination, we used the normalization operator $\mathcal{N}_1(.)$ (see Chapter 2.4.5).

Each scene considered in the experiments (Figure 3.5) is represented by its color (left) and its depth image (right). Under each feature map, the corresponding conspicuity map is represented. The two conspicuity maps are then combined, according to Equation 3.6, into the saliency map, which is represented at the bottom of each figure.

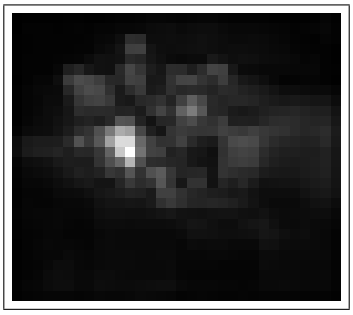
In the first scene (Figure 3.5), one attention spot is detected that stems from color contrast, whereas the depth feature gives rise to numerous equally conspicuous locations. Thus, the saliency map is strongly influenced by the color conspicuity map. Indeed, a kind of competition for saliency takes place between the two features. This competition is simulated by the normalization strategy $\mathcal{N}_1(.)$ used to combine the two conspicuity maps. As already mentioned in Chapter 2.4.5, $\mathcal{N}_1(.)$ promotes conspicuity maps in which a small number of strong peaks of activity is present, whereas maps that contain numerous comparable peak responses are suppressed. Thus, the large contribution of the color conspicuity map to the final saliency map is comprehensible.

In the second scene (Figure 3.6), exactly the opposite behavior occurs. The depth feature stimulates the detection of a single conspicuous location, whereas the color feature produces numerous comparable spots. Thus, the normalization operator $\mathcal{N}_1(.)$ promotes the depth conspicuity map and inhibits the color-based one. This explains the larger influence of the depth conspicuity map on the final saliency map.

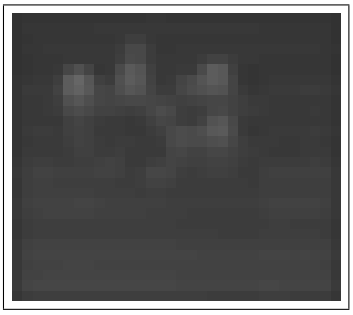
The third and the fourth examples (Figures 3.7 and 3.8) illustrate another

(a) Color image

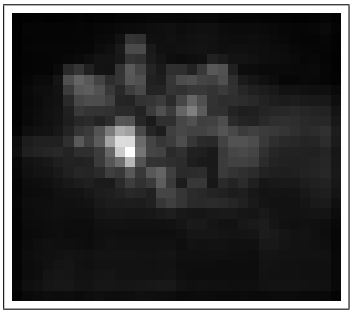
(b) Depth image



(c) Color conspicuity map

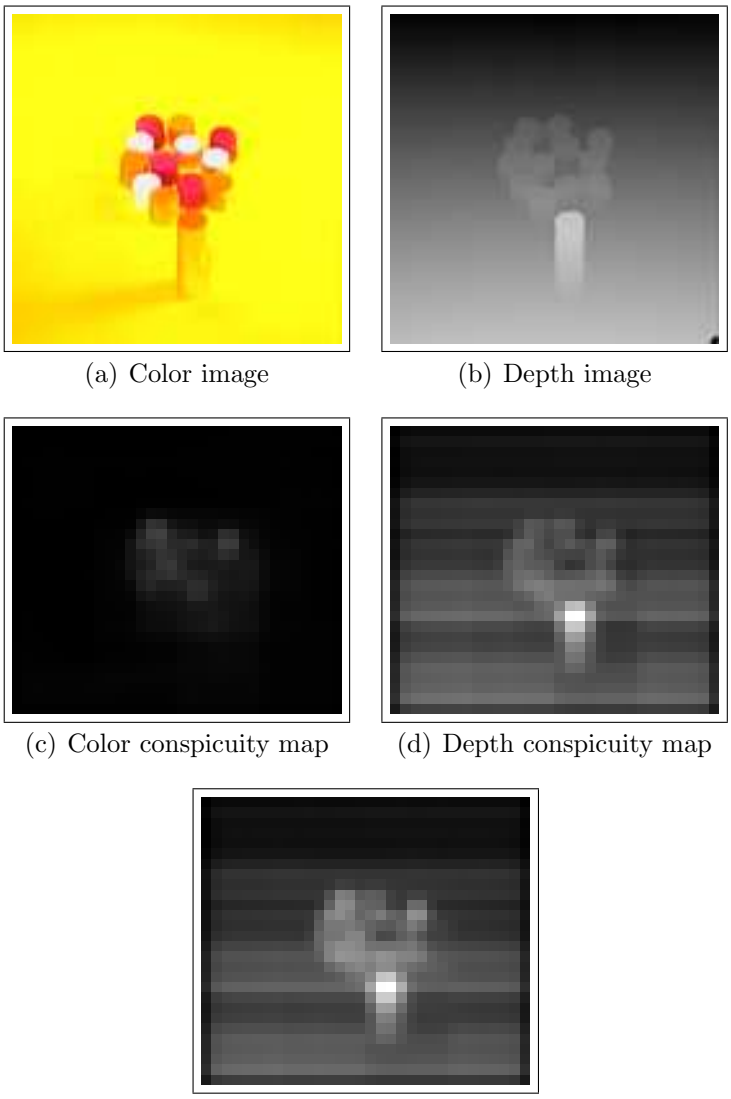


(d) Depth conspicuity map



(e) Saliency map

Figure 3.5: Results 1.

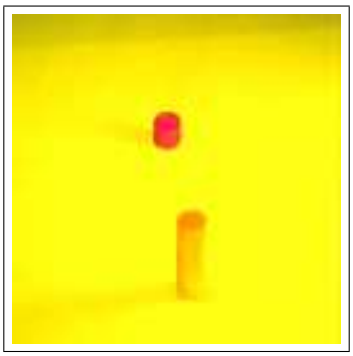


(a) Color image (b) Depth image

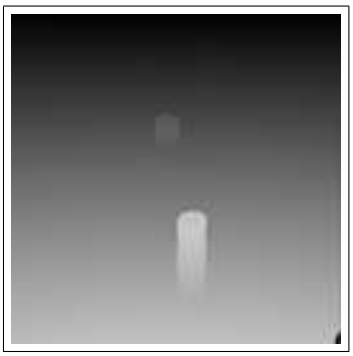
(c) Color conspicuity map (d) Depth conspicuity map

(e) Saliency map

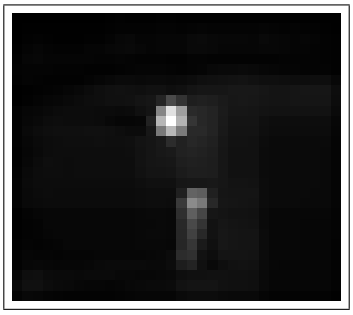
Figure 3.6: Results 2.



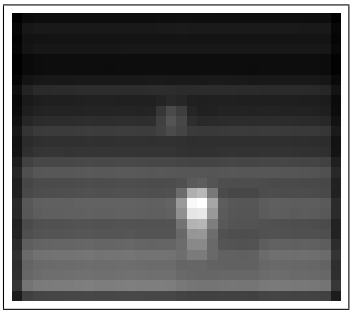
(a) Color image



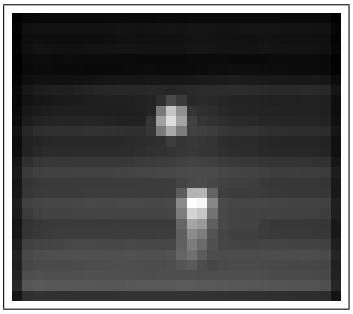
(b) Depth image



(c) Color conspicuity map

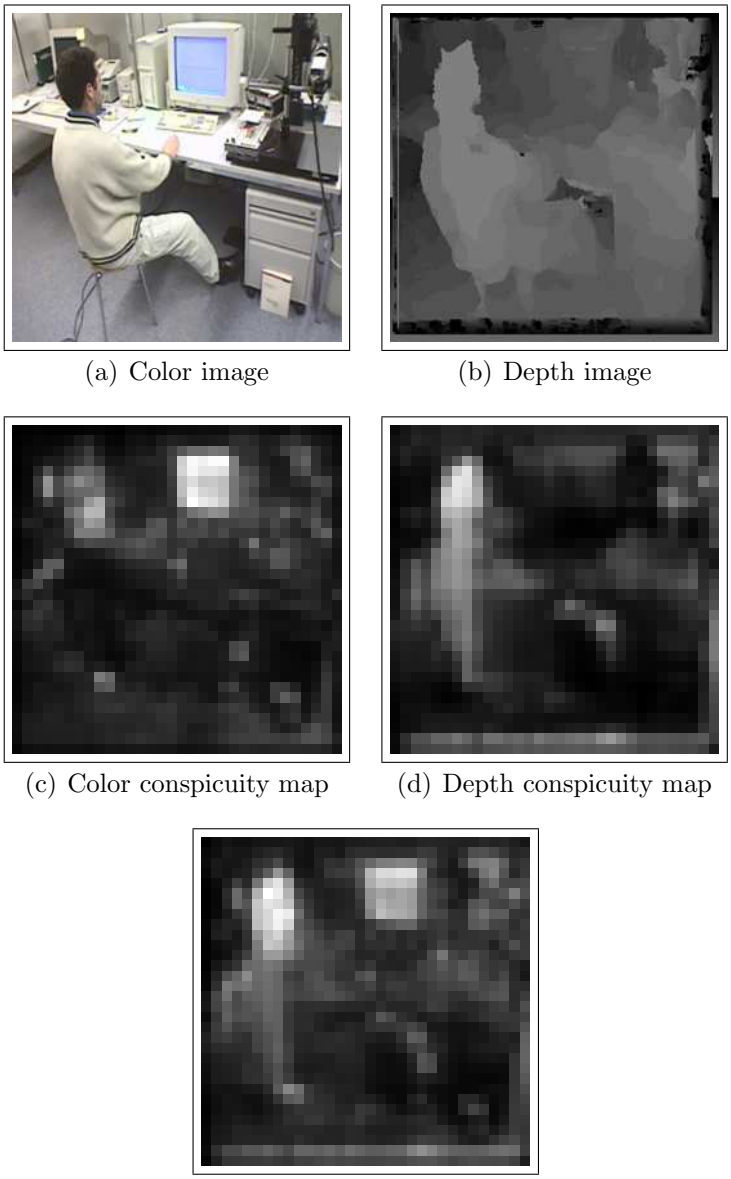


(d) Depth conspicuity map



(e) Saliency map

Figure 3.7: Results 3.



(a) Color image (b) Depth image

(c) Color conspicuity map (d) Depth conspicuity map

(e) Saliency map

Figure 3.8: Results 4.

characteristic of the proposed model, namely the cooperation between features to detect salient locations. In both examples, each of the two conspicuity maps contains a dominant conspicuous peak, which explains the promotion of both maps by the normalization operator $\mathcal{N}_1(.)$. Thus, the final saliency map is influenced by both maps and contains salient locations originating from both features.

By showing the clear usefulness of depth in scene analysis and its successful integration in a two-channel competitive task of visual attention, these experiments validate the depth enhanced computational model of attention. The effectiveness of channel competition is considered a key element for approaching further applications involving a larger number of cues.

3.3 Model of Dynamic Visual Attention

Motion is of fundamental importance in biological vision systems. The temporal aspect is highly relevant to the visual attention mechanism, since the rapid detection of moving objects, which may be preys or enemies, is often essential for the survival of species [161, 160, 58].

The importance of motion in computer vision cannot be understated, since a multitude of approaches rely on this cue to solve numerous computer vision problems. Typical applications where the motion cue plays a crucial role are structure from motion [12, 40], figure/ground segmentation [19, 30] and object tracking [107, 57].

Given the importance of motion, dynamic scene features must be considered in the early steps of computer vision, in particular in the pre-attentive stage. In a recent work [153] Tsotsos *et al.* confirmed the relevance of motion in the computational modelling of visual attention.

Regarding motion, we aim here at extending the saliency-based model of visual attention to consider also dynamic features, which gave rise to a model of dynamic visual attention [116]. The basic idea is to compute a conspicuity map related to motion which will be integrated with static conspicuity maps to compute the final saliency map, under the constraint that the motion computation method should be simple and fast. A multiscale gradient-based optical flow technique is used to compute the required dynamic conspicuity map.

3.3.1 Optical Flow Computation techniques

In an image, each pixel corresponds to the intensity value obtained by the projection of an object in 3-D space onto the image plane. When the objects move, their corresponding projections also change position in the image plane. Optical flow refers to the vector field that shows the direction and magnitude of these intensity changes from one image to the other, assuming that intensity (or color) of objects is conserved during displacement.

The computation of optical flow has been intensively investigated during the last decades, so that numerous approaches have been proposed to solve this problem. These approaches can be roughly classified into three main classes [11].

- **Differential techniques** [66, 101, 90, 154]. Also known as gradient-based techniques, they estimate optical flow from derivatives of image intensity over space and time.
- **Region-based matching techniques** [6, 138, 27]. They operate by matching small regions of image intensity or specific features from one frame to the next. The matching criterion is usually a least square or normalized correlation measure.
- **Energy-based and phase-based techniques** [62, 54]. These techniques rely on spatio-temporally oriented filters, and are typically derived by considering the motion problem in the Fourier domain.

In a complete work Barron *et al.* presented an empirical comparison of nine different optical flow estimation methods [11]. They concluded that the gradient-based methods provide the most accurate and the most dense measurements.

This technique, like the other two ones, fails however to correctly estimate or even detect large displacements in an image sequence. This drawback makes it necessary to extend the gradient-based optical flow estimation technique to a hierarchical form in order to cover a larger range of displacement scales. The hierarchical solution is adopted in this work.

In the following, the gradient-based optical flow estimation technique is first described and then its extension to a hierarchical form is reported.

Principle of Gradient-Based Optical Flow Estimation

As mentioned above, the optical flow estimation is associated to the variation of image intensity. The main assumption, made by several methods is the "brightness conservation". A physical point (or object) is supposed to have the same brightness throughout the sequence, even while moving. This assumption can be expressed by Equation 3.7.

$$I(\mathbf{x}, t) \;=\; I(\mathbf{x} + \mathbf{v} \cdot \delta t,\; t + \delta t) \tag{3.7}$$

where $I(\mathbf{x}, t)$ is the intensity of the point $\mathbf{x}$ at time t and $\mathbf{v} = (u, v)$ is its motion vector.

From a Taylor expansion of Equation 3.7, the *gradient constraint equation* is derived according to Equation 3.8 (more details on this derivation are given in B.1).

$$\nabla I(\mathbf{x}, t) \;.\; \mathbf{v} + I_t(\mathbf{x}, t) = 0 \tag{3.8}$$

where ∇I is the spatial gradient and I_t is the temporal derivative of I.

It is obvious that this equation alone is not sufficient to estimate both components of the motion vector, because we have an equation in two unknowns. This is known as the aperture problem. Generally, additional assumptions are needed in order to obtain a well-determined system of equations that yields both components of the motion vector at each point of the image [90, 66]. Nevertheless, these methods resolve the aperture problem at cost of high computation time.

Using the *gradient constraint equation*, the normal component of the motion vector can, however, be directly computed. The normal component of the motion vector is defined as the vector component orthogonal to spatial contours, i.e. in the gradient direction. Let $\mathbf{v}_n = s\mathbf{n}$ be this vector component. The normal speed s and the normal direction $\mathbf{n}$ are given by Equation 3.9.

$$s(\mathbf{x},t) \;=\; \frac{-I_t(\mathbf{x},t)}{||\nabla I(\mathbf{x},t)||} \qquad \mathbf{n} \;=\; \frac{\nabla I(\mathbf{x},t)}{||\nabla I(\mathbf{x},t)||} \tag{3.9}$$

Hierarchical Gradient-based Optical Flow Estimation

The gradient-based optical flow estimation is based on a fundamental assumption which is the smoothness of motion. Thus, classical gradient-based methods do

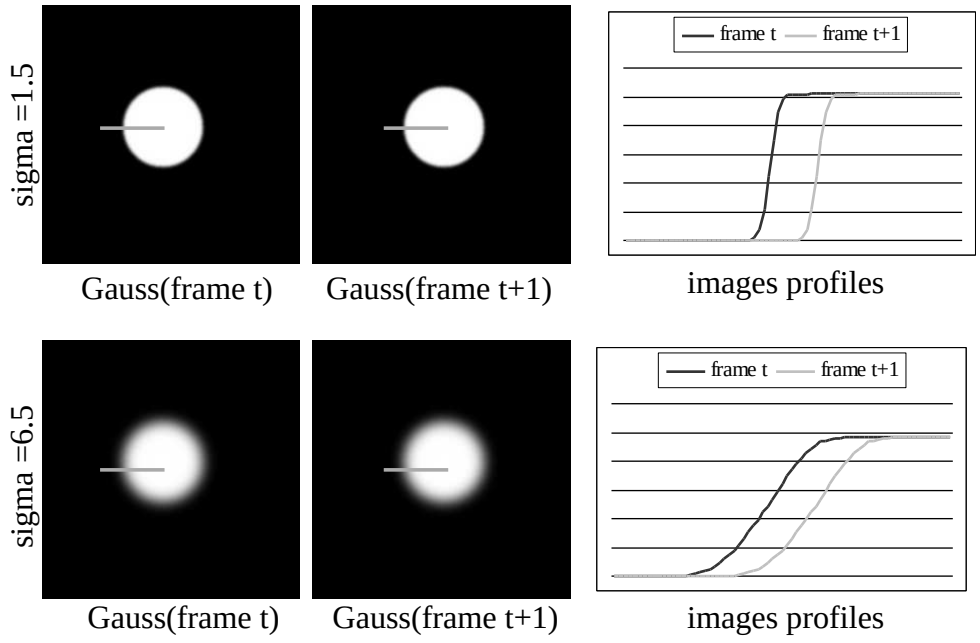


Figure 3.9: Large displacements and the multiscale concept. The first row represents two frames (t and $t+1$) of a moving disc smoothed with a small gaussian filter ($\sigma = 1.5$). The second row shows how the same edges perfectly recover when they are smoothed by a larger gaussian filter ($\sigma = 6.5$). It is obvious that motion of the disc can be estimated only at $\sigma = 6.5$.

not allow proper handling of large displacements.

Indeed, the estimation of optical flow of a moving physical edge is possible only if its images in two successive frames spatially overlap, as shown on Figure 3.9 (second row). When the displacement of the edge is too large, then the overlap condition is no more valid and consequently, the motion of that edge can not be computed.

Some previous works have proposed interesting approaches to overcome this drawback of gradient-based techniques by extending existing algorithms to deal also with large displacements. Very promising results have been obtained by introducing the multiscale concept to the optical flow estimation techniques [80, 13, 137, 136]. The basic idea is to detect large displacements at coarse scales and the small displacements at fine scales. Figure 3.9 illustrates the basic idea of this concept by showing how the multiscale technique resolves the problem of large displacements.

3.3.2 Dynamic Conspicuity Map

This section aims at computing a conspicuity map related to motion for each image of a sequence.

Based on the arguments given above, we decided to use, in this work, a multiscale gradient-based method in order to compute motion from image sequences. Furthermore, the absolute value of the normal speed s given by Equation 3.9 (hereafter the velocity) is used as indicator of motion. Surely this component does not correspond to the real motion vector (due to the aperture problem), it provides, however, a good approximation of scene motion. We believe that an approximative estimation about scene motion is sufficient in the pre-attentive stage. Accurate motion estimation can be achieved rather during the attentive stage by refining the rough estimation. The coarse-to-fine approach proposed by Simoncelli is a possible solution to the estimation refinement problem [137].

Given two successive frames, the implemented multiscale concept consists in building a gaussian resolution pyramid for each frame by progressively lowpass filtering and subsampling the original image (see Figure 3.10). Five resolutions are used in our implementation. For each of the five spatial resolutions a velocity map is computed using the gradient-based method (Equation 3.9). Thus, a velocity pyramid is obtained. Its fine scales detect small displacements, whereas its coarse scales detect large displacements.

From the computed velocity pyramid, a conspicuity map related to motion and which can be integrated into the model of visual attention should be now derived. This map has to represent really moving objects. Two scenarios can be conceived in this context. If we deal with moving cameras, one should detect locations that move differently from the rest of the scene (background). In this case, a conspicuity operator is essential to discriminate such objects. On the other hand, when using a stationary camera, the velocity map itself clearly discriminates

the moving objects of the scene. This work addresses the latter scenario. We derive a dynamic conspicuity map C_d by summing the five velocity maps at the coarsest resolution (which is the resolution of the conspicuity maps related to static features). The velocity maps at finer resolutions are subsampled to the

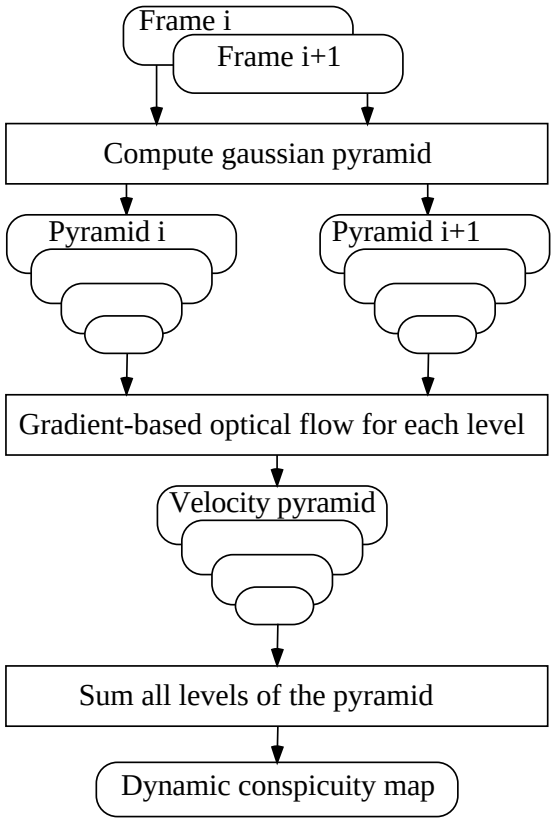


Figure 3.10: Computing of the dynamic conspicuity map. Given two successive frames of an image sequence, a gaussian pyramid is computed from each frame. Using a gradient-based technique, a velocity map is computed for each pair of images of the same level, giving rise to a velocity pyramid. Finally, all levels of the velocity pyramid are summed together into the dynamic conspicuity map.

desired resolution after a lowpass filtering. Figure 3.11 gives an example of the described dynamic conspicuity map from two frames of the well-known "Taxi Hamburg" sequence.

3.3.3 Combining Static and Dynamic Conspicuity Maps

This section describes the integration of the dynamic scene features into the saliency-based model which considered so far only static features. The basic idea is to combine, into a final saliency map, a conspicuity map C_s related to static

(a) frame 8 (b) frame 9 (c) Dynamic conspicuity map

Figure 3.11: "Taxi Hamburg": dynamic conspicuity map.

features and the dynamic conspicuity map C_d for each frame t of the image sequence. It is noteworthy that the dynamic conspicuity map is computed from two successive frames and that for each frame, a saliency map is computed independently from the saliency results of previous frames. A schematic description of the resulting model - the model of dynamic visual attention - is depicted on Figure 3.12.

We propose two different strategies to combine the two conspicuity maps C_s and C_d into the final saliency map.

Competition-Based Strategy

The first method consists in a pure data-driven competition between static and dynamic scene features. For each frame t the corresponding saliency map $S_1(\mathbf{x}, t)$ results from a weighted summation of the two conspicuity maps (Equation 3.10).

$$S_1(\mathbf{x}, t) = \mathcal{N}_1(C_s(\mathbf{x}, t)) + \mathcal{N}_1(C_d(\mathbf{x}, t)) \tag{3.10}$$

where $\mathcal{N}_1(.)$ is the contents-based global amplification normalization method presented in Chapter 2.4.5 and which promotes conspicuity maps in which a small number of strong peaks of activity is present. Maps that contain numerous comparable peak responses are demoted.

Motion-Conditioned Strategy

The second strategy of integration prioritizes the motion cue. This means that only moving objects can compete for saliency. In this case the corresponding saliency map $S_2(\mathbf{x}, t)$ is computed as follows:

$$S_2(\mathbf{x}, t) = \begin{cases} C_s(\mathbf{x}, t) & if\ C_d(\mathbf{x}, t) > \epsilon \\ 0 & otherwise \end{cases} \tag{3.11}$$

This strategy can be useful in computer vision applications where only moving scene objects are of special interest, like in video surveillance.

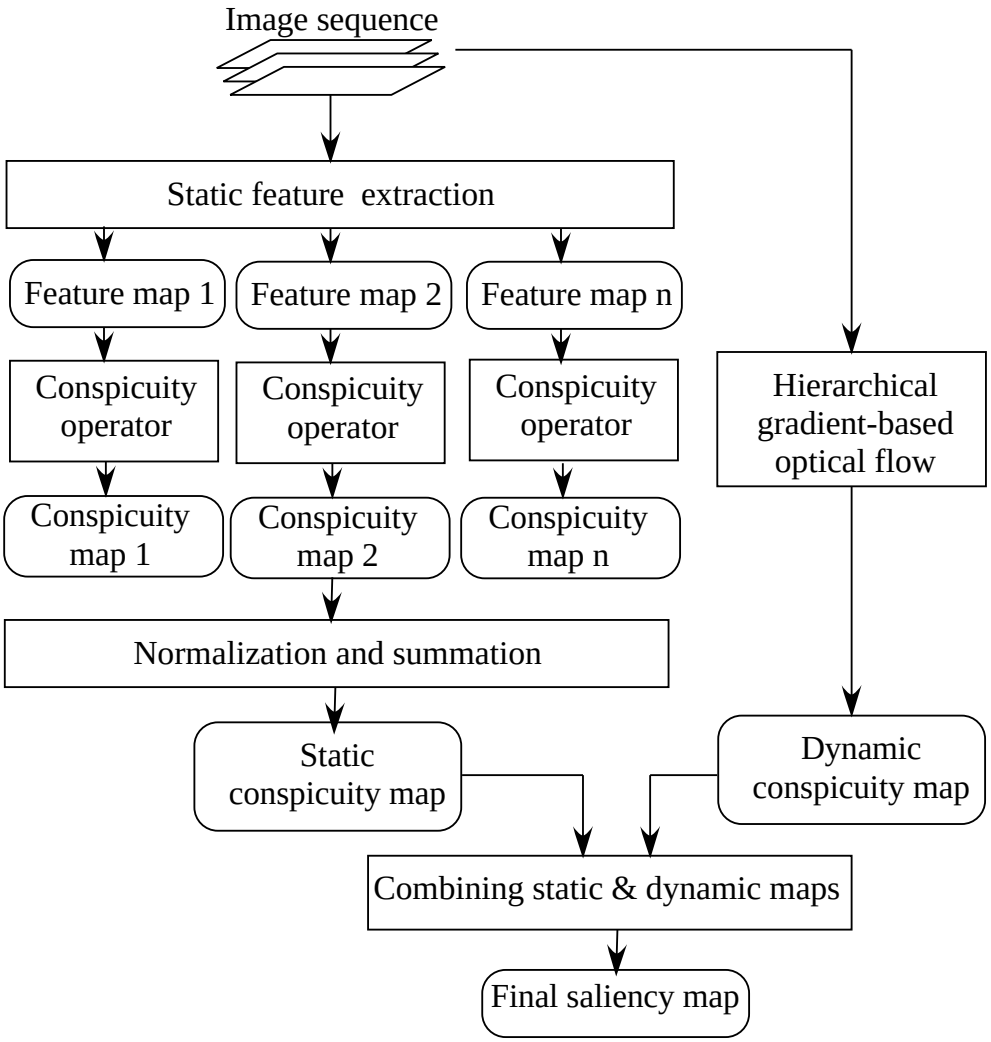


Figure 3.12: Model of dynamic visual attention.

Depending on the application nature, the appropriate integration strategy can be chosen. We introduce a parameter α that controls this choice according to Equation 3.12.

$$S = \alpha.S_1 + (1 - \alpha)S_2 \tag{3.12}$$

Thus, if $\alpha = 1$, then a competition between static and dynamic features takes place, whereas only moving objects compete for saliency, if $\alpha = 0$.

The most visually salient image locations (objects) are derived from the final saliency map S by applying a Winner-Take-All algorithm, as described in Chapter 2.4.4.

3.3.4 Results and Discussion

This section reports experiments carried out on real color image sequences in order to assess the presented model of dynamic visual attention.

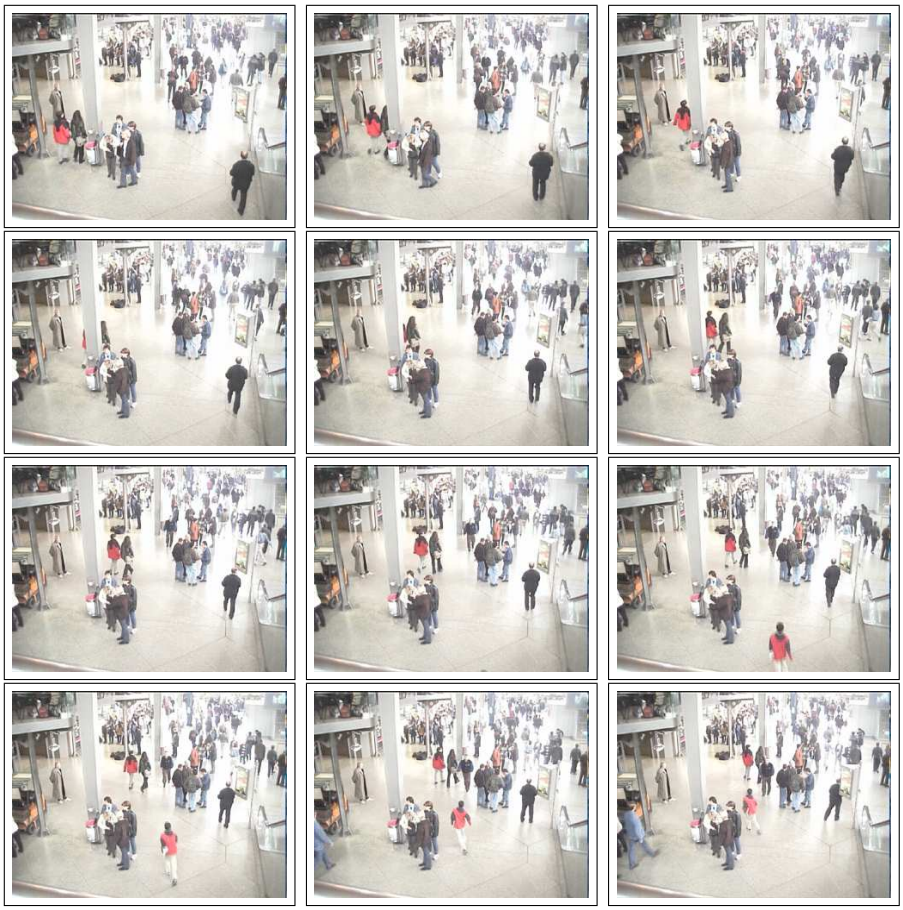


Figure 3.13: "Munich Train Station" sequence. Each fifth frame is represented.

The presented test image sequence has been acquired at Munich train station using a commercial digital video camera (Figure 3.13). In these experiments, we compute the static conspicuity map from the two chromatic features $(R-G)$ and $(B-Y)$.

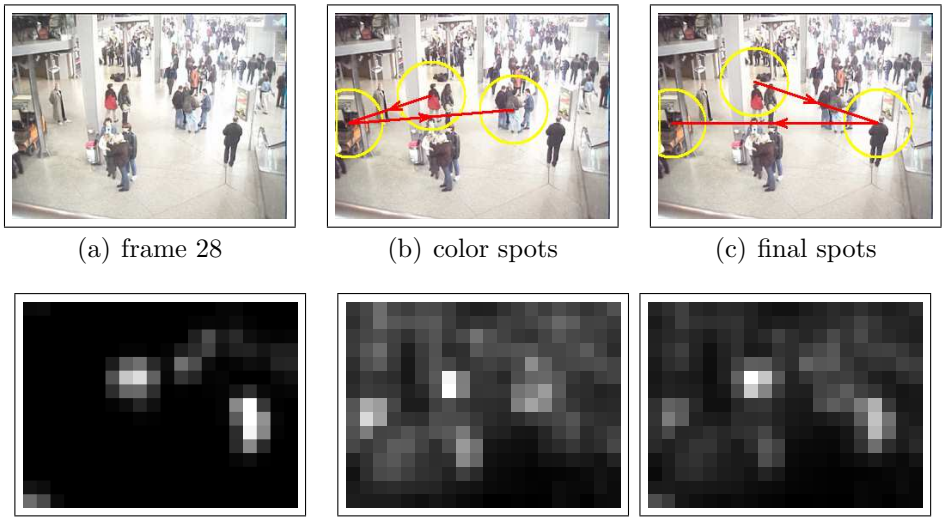
In a first experiment (Figure 3.14), a dynamic conspicuity map (Figure 3.14 (d)) as well as a conspicuity map related to chromatic features (Figure 3.14 (e)) have been computed for each frame. Both conspicuity maps are then integrated into the final saliency map (Figure 3.14 (f)) using the competition-based integration strategy (Equation 3.10). For comparison purposes, two sets of 3-spot-of-attention are computed. The first set (Figure 3.14 (b)) is derived from the color conspicuity map, whereas the second one (Figure 3.14 (c)) is computed from the final saliency map (motion + color). The spots are linked with arrows in the direction of decreasing saliency. Considering only chromatic features the attention mechanism detects the walking woman (in red), then a stationary wagon, and finally a group of standing people. Considering also motion, the walking woman still holding the highest value of saliency (due to her salient color and motion). The second salient position is no more attributed to a stationary object, but to the walking man in black (right image border), who won this position, exclusively, due to his motion. The third position is, however, attributed to the stationary wagon.

To summarize, this version of the attention mechanism creates a competition between static and dynamic features for saliency. A cooperation between the two feature classes can, however, take place allowing the discrimination of scene parts which have salient static and dynamic features (the case of the walking woman in red).

The second experiment presented here (Figure 3.15) deals with the same scene. The motion-conditioned integration strategy (Equation 3.11) is now used to compute the final saliency map S_2 from C_s and C_d. Unlike the first experiment, no stationary objects figure within the detected salient locations. Furthermore, the chromatic conspicuity map determines the order of saliency of the moving scene parts.

The third experiment (Figure 3.16) gives a further example of the motion-conditioned integration strategy. Only moving scene parts are detected, and they are ordered according to their color conspicuousness.

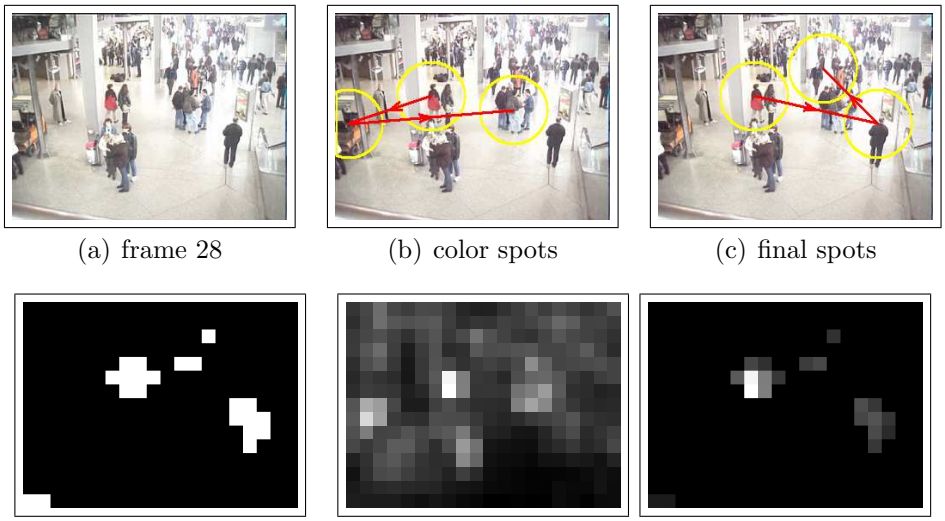
To conclude, the experiments clearly show the usefulness of motion in the attention mechanism. For both integration strategies the dynamic conspicuity map strongly influences the behavior of the model by increasing the overall saliency of moving scene objects.



(a) frame 28 (b) color spots (c) final spots

(d) C_d (e) C_s (f) S_1

Figure 3.14: Integration of motion and color: competition-based strategy.



(a) frame 28 (b) color spots (c) final spots

(d) C_d thresholded (e) C_s (f) S_2

Figure 3.15: Motion and color Integration: motion-conditioned strategy.

Figure 3.16: Tracking of the most salient moving objects.

3.4 Chapter Summary

This chapter has presented two major extensions of the saliency-based model of visual attention. The first part of this chapter has discussed the integration of depth-related features into the model of visual attention so that the visual attention algorithm can operate on 3D scenes. Therefore three depth-based features has been considered. Preliminary results tended to show that the depth is highly significant for visual attention, whereas the assessment of the usefulness of mean curvature and of the depth gradient is still open. Then, a combination method of 2D and 3D feature into a unique saliency map has been described. The experiments carried out with synthetic and real 3D scenes clearly showed the usefulness and the complementarity of the depth feature for the visual attention model.

The second part of the current chapter has reported a model of dynamic visual attention, that is the detection of salient objects in a dynamic scene. In addition to a static conspicuity map, computed from static scene features like color and intensity, the proposed model computes a dynamic conspicuity map which highlights the moving parts of the scene. Two strategies have been then proposed to combine the static and the dynamic maps. The first strategy consists in a pure data-driven competition, whereas the second one prioritizes the motion cue and thus allows only moving objects to be identified as salient parts of the scene. For both strategies, the high relevance of the motion cue for visual attention has been shown by means of experiments carried out with real image sequences. Indeed, the dynamic features strongly influence the behavior of the attention model by considerably increasing the saliency of moving scene objects.

Chapter 4

Empirical Validation of the Visual Attention Model

4.1 Chapter Introduction

In human vision, it is believed that visual attention is intimately linked to the eye movements and that the fixation points correspond to the location of the salient image locations [165]. Thanks to the availability of sophisticated eye tracking technologies, several recent works have confirmed this link between visual attention and eye movements [82, 133, 126]. Hoffman et *al.* suggested in [65] that saccades to a location in space are preceded by a shift of visual attention to that location. Using visual search tasks, Findlay and Gilchrist concluded that when the eyes are free to move, no additional covert attentional scanning occurs, and most search tasks will be served better with overt eye scanning [51, 52]. Maioli et *al.* go further to say that "There is no reason to postulate the occurrence of shifts of visuospatial attention, other than those associated with the execution of saccadic eye movements" [91].

Thus, eye movement recording is a suitable means for studying the temporal and spatial deployment of attention in any situation.

In computer vision, the saliency-based model of visual attention is commonly accepted and used in the field, despite the fact that only few works have dealt with the assessment of its biological grounding [121].

This chapter deals with the empirical validation of the saliency-based model of visual attention by comparing its performance with the human attention [157, 119]. The validation method quantitatively evaluates the contribution of color to visual attention controlling and it assesses the plausibility of different working modes of the model like the linear versus non-linear combination ($\mathcal{N}_1(.)$ and $\mathcal{N}_2(.)$) of visual cues into the final attention map.

4.1.1 Chapter Outline

The remainder of the current chapter is organized as follows. Section 4.2 presents the general approach of empirically assessing the plausibility of the computational model of visual attention. Section 4.3 reports the results obtained from the empirical experiments and interpret them regarding different aspects. This section assesses, for instance, the plausibility of the $\mathcal{N}_1(.)$-based model and the $\mathcal{N}_2(.)$-based one by measuring their respective correlations with human visual attention. It also evaluates, quantitatively, the contribution of color to visual attention. Finally, Section 4.4 summarizes and concludes the chapter.

4.2 Overview of the Method

The basic idea of our validation method is to compare the computational map of attention produced by the model of visual attention with another map derived from eye movement experiments, i.e. the human attention map. Practically, a computational attention map is computed for a given image and considering some features. The same image is presented to human subjects whose eye movements are recorded, providing information about the spatial location of the sequentially attended image locations and the duration of each fixation. A human attention map is then derived, under the assumption that this human attention map is an integral of single impulses located at the positions of the successive fixation points. Objective comparison criteria have been established in order to measure the similarity of both maps.

Algorithm 4.1 illustrates the main steps of the proposed validation method.

Algorithm 4.1 Empirical method for model validation

(1) Compute the computational map of attention from an image
(2) Record subjects eye movement for the same image
(3) Compute human attention map
(4) Compare computational and human attention maps

The different steps of the approach are detailed in the following sections.

4.2.1 Computational Map of Attention

The computational attention map is computed by means of the saliency-based model of visual attention. This step will be described for each version of the computational model to be assessed (Section 4.3.1 and 4.3.2).

4.2.2 Human Map of Attention

As discussed above, the deployment of visual attention in humans is intimately linked to their eye movements. Under the assumption that attention is guided by the saliency of the different image locations, the recorded fixation locations can serve to plot a saliency distribution map. The computation of the human attention map is achieved in two steps. First, eye movements of several human subjects are recorded while they are viewing a given image. Second, these measurements are transformed into a human map of attention.

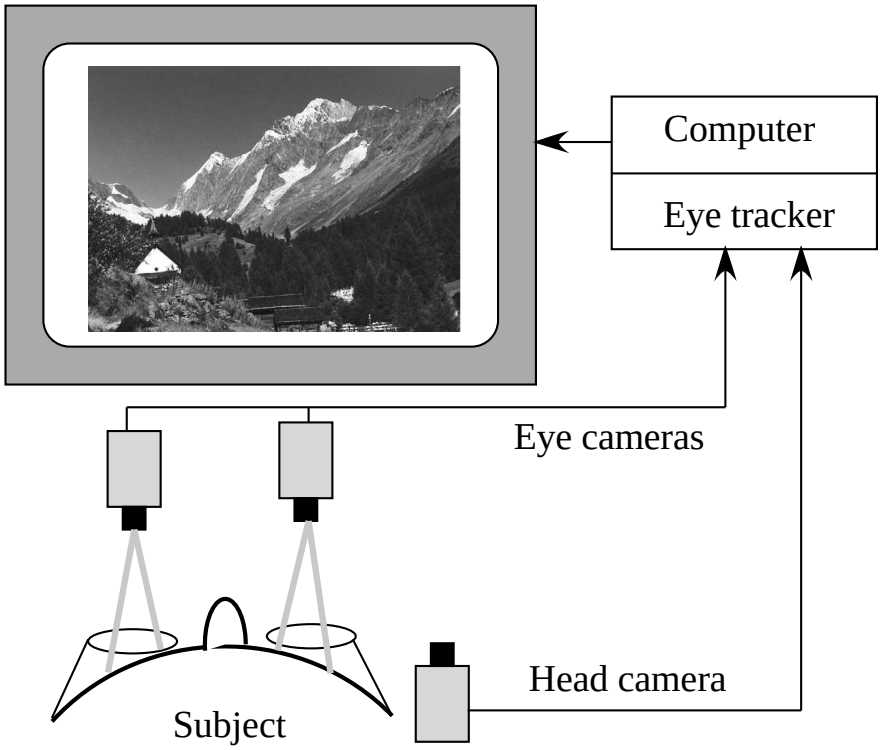


Figure 4.1: Principle of eye movements recording.

Eye Movement Recording

Eye position was recorded with a video-based tracking system (EyeLink^{TM}, SensoMotoric Instruments^{TM} GmbH, Teltow/Berlin) at the Perception and Eye Movements Laboratory [122]. This system consists of a headset with a pair of infrared cameras tracking the eyes (*eye cameras*), and a third camera (*head camera*) monitoring the screen position in order to compensate for any head movements (see

subj	image	fix	time(ms)	duration(ms)	x	y
407	10	1	407	240	205	382
407	10	2	679	228	137	440
407	10	3	935	364	139	353
407	10	4	1319	236	144	398
407	10	5	1623	252	545	523
407	10	6	1911	176	432	538
407	10	7	2119	188	604	581
407	10	8	2391	288	454	143
407	10	9	2707	352	492	87
407	10	10	3095	280	426	230
407	10	11	3415	368	595	202

Figure 4.2: Example of eye movement data for one human subject.

Figure 4.1). Once the subject placed in front of a display for stimulus presentation, the position of his pupils is derived from the *eye cameras* using specialized image processing hard- and software. Combining the pupil and head positions, the system computes the gaze position at a rate of 250 Hz and with a gaze-position accuracy relative to the stimulus position of $0.5^o - 1.0^o$, largely dependent on subjects fixation accuracy during calibration.

Indeed, each experimental block was preceded by two 3x3 point grid calibration sequences, which the subjects were required to track. The first calibration scheme is part of the EyeLink system and allows for on-line detection of non-linearities and correction of changes of headset position. The second calibration procedure is supported by an interactive software in order to obtain a linearized record of eye movement data.

The eye tracking data are parsed for fixations and saccades in real-time, using parsing parameters proven to be useful for cognitive research thanks to the reduction of detected microsaccades and short fixations ($<$ 100 ms). Remaining saccades with amplitudes less than 20 pixels (0.75^o visual angel) as well as fixations shorter than 120 ms were discarded afterwards.

Figure 4.2 illustrates an example of eye movement data recorded for one human subject while looking at an image during about 5 seconds.

Computation of the Human Attention Map

This section aims at computing a human attention map based on the fixation data collected in the experiments with human subjects. The basic idea is that this human attention map is an integral of single impulses located at the positions

of the successive fixation points. Practically, each fixated location gives rise to a normally (gaussian) distributed activity. The width (σ) of the activity patch can be set by the user and practically chosen to approximate the size of the fovea. We also introduced a parameter α that tunes the contribution of the fixation duration to the gaussian amplitude. On one hand, if $\alpha = 0$, the amplitude is the same for all fixations regardless of their duration. On the other hand, if $\alpha = 1$, the amplitude is proportional to the fixation duration.

The procedure to compute the human map of attention is described in Algorithm 4.2.

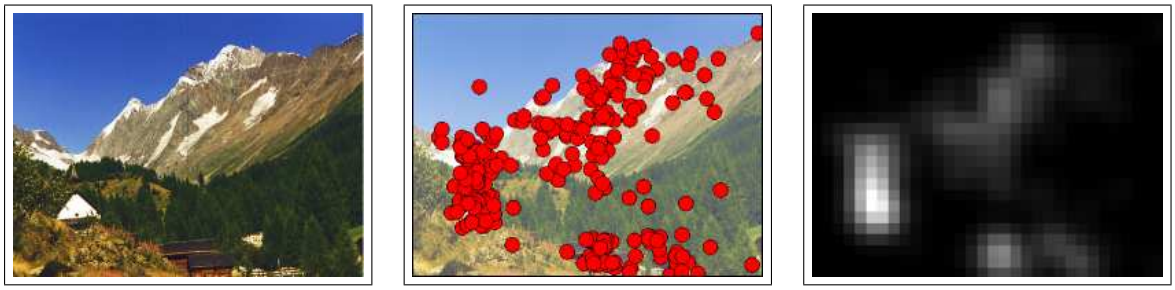
Algorithm 4.2 From fixation points to human attention map

Color image I ($w \times h$)
H : human attention map (output)
N eye fixations $F_i(x, y, t)$ ((x, y) are spatial coordinates and t is the duration of the fixation)
σ the standard deviation of the gaussian patch (FOVEA)
$\alpha \in [0..1]$ tunes the contribution of fixation duration to human attention
$Img1 = Img2 = 0$
$i = 1$
while $i \leq N$ **do**
  $(x, y) = Coord(F_i)$
  $t = Duration(F_i)$
  for $k = 0 \,..\, h - 1$ **do**
    for $l = 0 \,..\, w - 1$ **do**
      $Img1(k, l) = (\alpha.t + (1 - \alpha)).\exp(-\frac{(x-l)^2+(y-k)^2}{\sigma^2})$
    end for
  end for
  $Img2 = Img2 + Img1$
  $Img1 = 0$
end while
$Normalize(Img2, 0, 255)$
$H = subSample(Img2)$

Figure 4.3 gives an example of a human map of attention from eye movement data. In this example, the eye fixations have been recorded from 20 human subjects.

4.2.3 Comparison Metrics

The idea is to compare the computational attention map computed by the model of visual attention and the human attention map derived from human eye movement recordings. For this purpose two similarity measures have been used; cor-



(a) Original image (b) Recorded fixations (c) Human map of attention

Figure 4.3: Fixation to saliency: Example.

relation coefficient ρ and the relative fixation-to-chance distance Φ proposed in [121].

Correlation coefficient

Let $H(x)$ and $S(x)$ be the human and the computational maps respectively. The correlation coefficient ρ of the two maps is computed according to Equation 4.1.

$$\rho = \frac{\sum_x [(H(x) - \mu_h).(S(x) - \mu_s)]}{\sqrt{\sum_x (H(x) - \mu_h)^2 . \sum_x (S(x) - \mu_s)^2}} \tag{4.1}$$

where μ_h and μ_s are the mean values of the two maps $H(x)$ and $S(x)$ respectively.

Relative Fixation-to-Chance Distance

Unlike the correlation coefficient, the fixation-to-chance distance does not require the human map of attention, but the raw fixation data. The basic idea behind this method is to compare sample values extracted from the computational saliency map at locations determined by human fixations with sample values of the same saliency map extracted randomly [121].

Indeed, if the computation and the human behaviors correlate, then the computational saliency values at human fixation locations should be higher than the randomly selected saliency values. Let S be the computational saliency map computed from the input image and let $\mathcal{F}$ be the set of the eye fixations recorded for N human subjects and considering the k first fixations for each subject. In order to qualify the correlation between the computational saliency and the human fixations, the relative fixation-to-chance distance is computed as follows. First, the distribution $P_r(s)$ of the mean of $(N \cdot k)$ randomly selected saliency values s_r is computed (gaussian distribution on Figure 4.4) and its mean value s_μ is

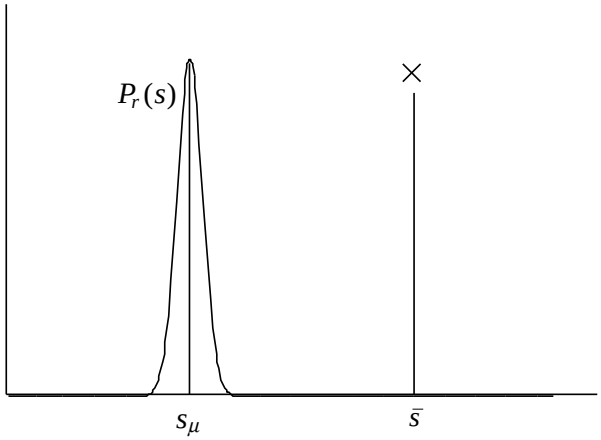


Figure 4.4: Relative fixation-to-chance distance.

determined in accordance with Equation 4.2.

$$s_\mu = \sum_{s=0}^{s_{max}} s \cdot P_r(s) \tag{4.2}$$

where s_{max} is the maximum value of S.

Second, for each recorded fixation $f_{i,j}$ in $\mathcal{F}$ (with $1 \leq i \leq k$ and $1 \leq j \leq N$) the corresponding spatial coordinates (x_{ij}, y_{ij}) are determined. Third, the mean saliency value $\overline{s}$ of the fixation locations is computed in accordance with Equation 4.3.

$$\overline{s} = \frac{1}{card(\mathcal{F})} \sum_{i=1}^{N} \sum_{j=1}^{k} S(x_{ij}, y_{ij}) \tag{4.3}$$

Finally the fixation-to-distance Φ is computed according to Equation 4.4.

$$\Phi = \frac{\overline{s} - s_\mu}{\overline{s}} \tag{4.4}$$

Note that if Φ is positive, then the fixation-guided saliency values are higher than the randomly selected saliency values, which indicates a correlation between the human and the computational map of attention. On the other hand, if Φ is negative, then the computational and the human attention maps are anti-correlated.

4.3 Experiments and Discussion

In this section we report some experiments conducted in order to asses the similarity between the human and the computational maps of attention. The above described comparison methodology is used for this purpose.

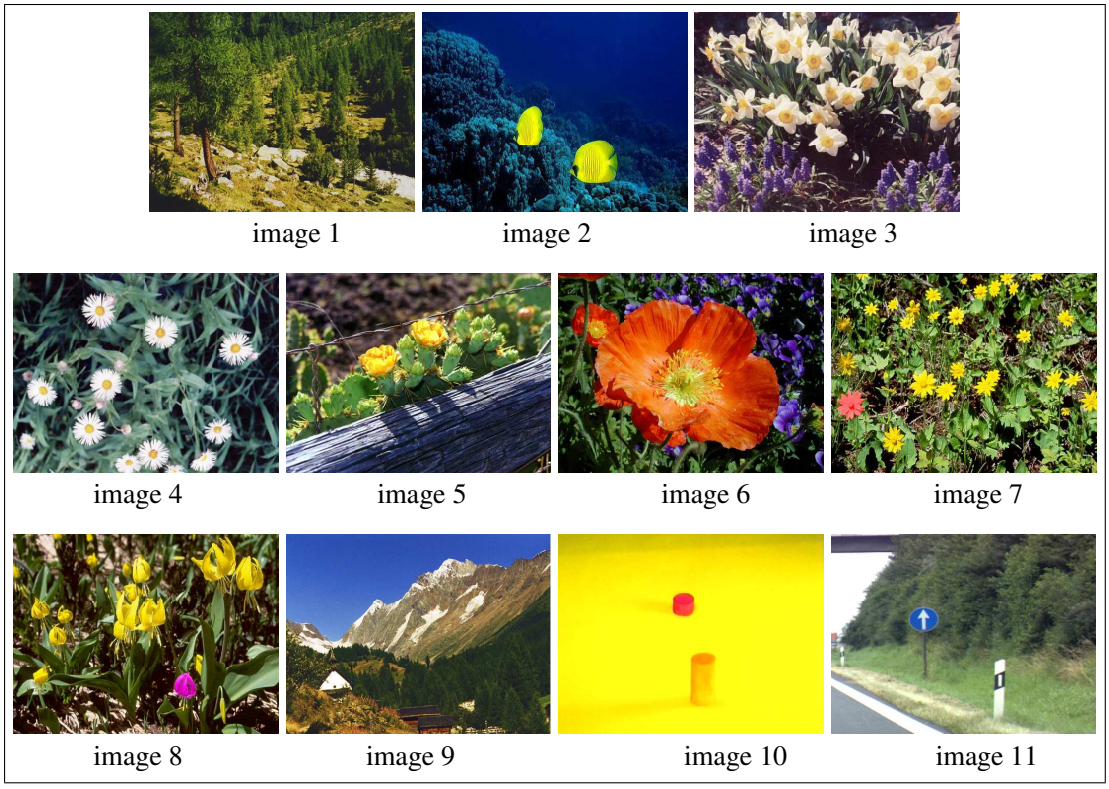


Figure 4.5: Images used in the experiments.

The eye recording experiments were conducted with 7 subjects between 24 and 34 years, 6 female and 1 male. All of them have normal or corrected-to-normal visual acuity as well as normal color vision. The images (see Figure 4.5) were presented to the subjects with a resolution of 800×600 pixels on a 19" monitor, placed at a distance of 70 cm from the subject, which results in a visual angle of approximatively 29×22^o. Each image was presented for 5 seconds, during which the eye movements were recorded. The instruction given to the subjects was "just look at the image".

From the recorded eye movement data, we compute human maps of attention using the following parameter values:

- $\sigma = 37.0$ for all maps.
- $\alpha = 0$. That means that all fixations have the same importance, regardless of their duration.

Two versions of the saliency-based model of visual attention are assessed in this section. The first version uses the global amplification normalization strategy ($\mathcal{N}_1(.)$) to normalize the different conspicuity maps before integrating them into the final saliency map, whereas the second version uses the nonlinear normalization strategy ($\mathcal{N}_2(.)$) for the same purpose. In both versions of the model three visual cues are considered: the color cue which is represented by two features $(R-G)$ and $(B-Y)$; the intensity cue which contains a single feature I; and the orientation cue which is composed of four orientation features $(0^o, 45^o, 90^o, 135^o)$.

4.3.1 Validation of the $\mathcal{N}_1(.)$-based Model

This section reports the assessment of the $\mathcal{N}_1(.)$-based version of the visual attention model.

image	1	2	3	4	5	6	7	8	9	10	11	mean
ρ (%)	50.5	46.4	33.5	26.1	39.2	38.5	29.4	38.7	52.5	65.8	23.6	42.0
Φ (%)	21.7	75.9	17.0	36.5	43.0	30.8	41.1	35.1	54.7	84.8	36.9	43.4

Table 4.1: Similarity measures between the human and the computational attention map for the $\mathcal{N}_1(.)$-based model of visual attention, using correlation coefficient (ρ) and relative fixation-to-chance distance (Φ).

The similarity measures between the human and the computational attention map for this version of the attention model are represented in Table 4.1.

It can be observed that the correlation coefficient ρ varies between 23% and 66% ($\rho \in [23..65\%]$) and that its mean value exceeds 40%. The relative fixation-to-chance distance Φ lies in an interval [17..85%] with a mean value of over 43%.

The positive scores of both measures with all images on one hand, and the relative high values of their respective mean values speak for the similarity between the human and the computational map of attention.

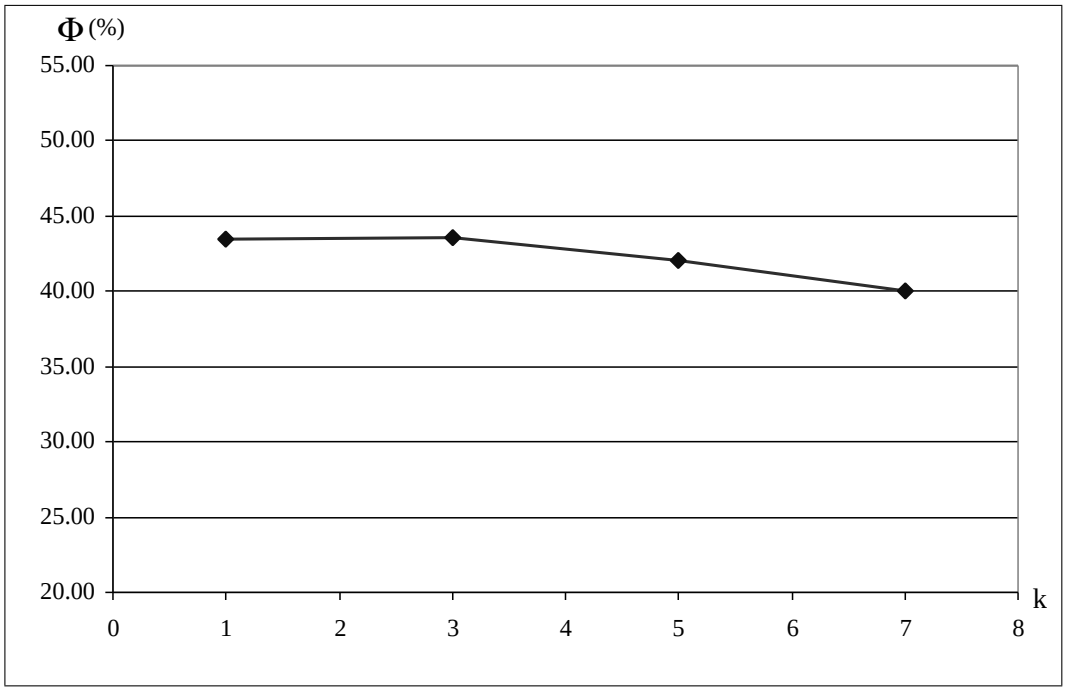


Figure 4.6: Impact of presentation time on Φ.

The impact of the presentation time of the images to the human subjects (which is expressed by the number of fixations k) on the similarity between the human and the computational attention map is also examined. The relationship between the mean fixation-to-chance distance Φ (for all subjects and all images) and the number of the k first fixations is illustrated on Figure 4.6. It can be remarked that the similarity between the human and the computational map decreases with k, which is consistent with the assumption of the slow onset of top-down attentional effects [121]. These observations are in accordance with the bottom-up nature of our visual attention model.

4.3.2 Validation of the $\mathcal{N}_2(.)$-based Model

It has been pointed out in Section 2.4.5 that the $\mathcal{N}_1(.)$ normalization strategy has been considered for its computational simplicity rather than for its biological plausibility. $\mathcal{N}_2(.)$ normalization has been, however, qualified as biologically plausible [69]. The aim of this section is to quantitatively measure the plausibility of the $\mathcal{N}_2(.)$-based version of the visual attention model with the human visual attention and to compare it with the plausibility of the $\mathcal{N}_1(.)$-based version.

According to Equation 2.9, the $\mathcal{N}_2(.)$ normalization method is characterized by three main parameters, namely σ_{on} and σ_{off} of the $\mathcal{D}o\mathcal{G}$ filter and the number of iterations p. In our experiments where the conspicuity maps to be normalized have a typical size of 32×24 pixels, we set these parameters as follows:

- $\sigma_{on} = 1.0$
- $\sigma_{off} = 6 \cdot \sigma_{on} = 6.0$
- $p = 3$

image	1	2	3	4	5	6	7	8	9	10	11	mean
ρ (%)	27.7	63.0	19.6	25.6	54.9	52.7	39.7	56.9	41.8	66.5	41.8	44.6
Φ (%)	40.7	90.2	7.8	68.9	82.8	74.2	78.7	80.0	77.3	95.2	74.5	70.0

Table 4.2: Similarity measures between the human and the computational attention map for the $\mathcal{N}_2(.)$-based model of visual attention, using correlation coefficient (ρ) and relative fixation-to-chance distance (Φ).

The similarity measures for the comparison of the human and the $\mathcal{N}_2(.)$-based computational map of attention are resumed in Table 4.2. The table shows that ρ lies in an interval [19..67%] and its mean value amounts to 44%. It also shows that Φ varies between 8% and 95% ($\Phi \in [8..95\%]$) with a mean value of 70%. The same indexes as for the $\mathcal{N}_1(.)$-based model, namely the positive scores for all images and the even higher corresponding mean values, point to the similarity between the human and the $\mathcal{N}_2(.)$-based computational map of attention.

Figures 4.7 and 4.8 give a direct comparison of both versions of the model regarding their plausibility. In general, the $\mathcal{N}_2(.)$-based computational map of attention correlates more with the human attention map than the $\mathcal{N}_1(.)$-based computational map does, which confirms the theoretical thoughts stated in [69] about the plausibility of the nonlinear normalization strategy.

According to Figure 4.8, few images do not obey to the general constatation made above about the superiority of the $\mathcal{N}_2(.)$-based model over the $\mathcal{N}_1(.)$-based one. Note that an adaptation of the parameters σ_{on}, σ_{off} and p of the $\mathcal{N}_2(.)$ normalization method strongly enhances the correlation coefficients of the human and computational attention map for these images. In fact, these parameters represent a potential tuning mechanism to control the behavior of the computational model depending on image types.

4.3.3 Color Contribution to Visual Attention

This section aims at assessing the role of chromatic features in controlling visual attention. The approach used for this purpose consists in comparing the human

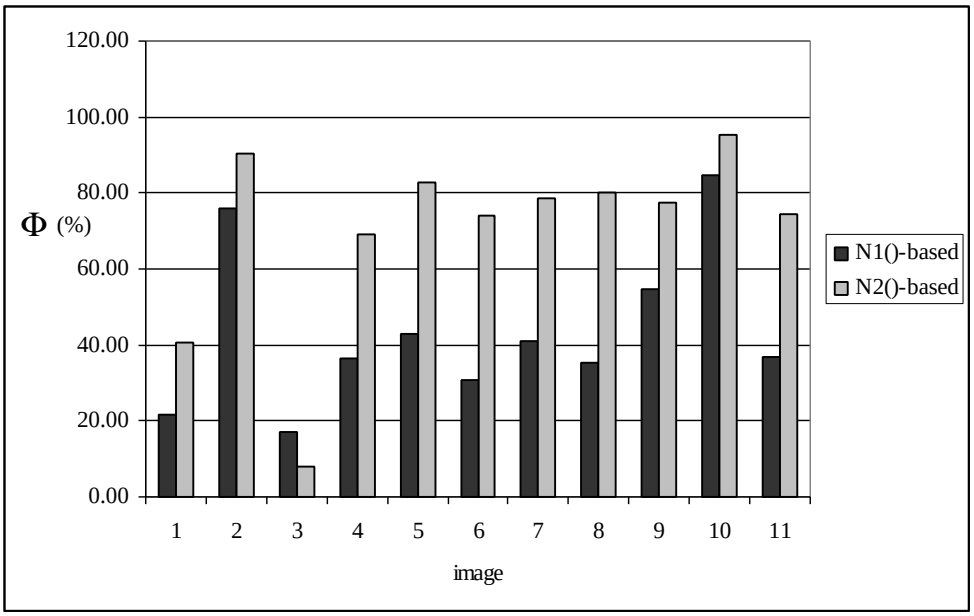


Figure 4.7: $\mathcal{N}_1(.)$-based model vs. $\mathcal{N}_2(.)$-based model: Fixation-to-chance distance Φ.

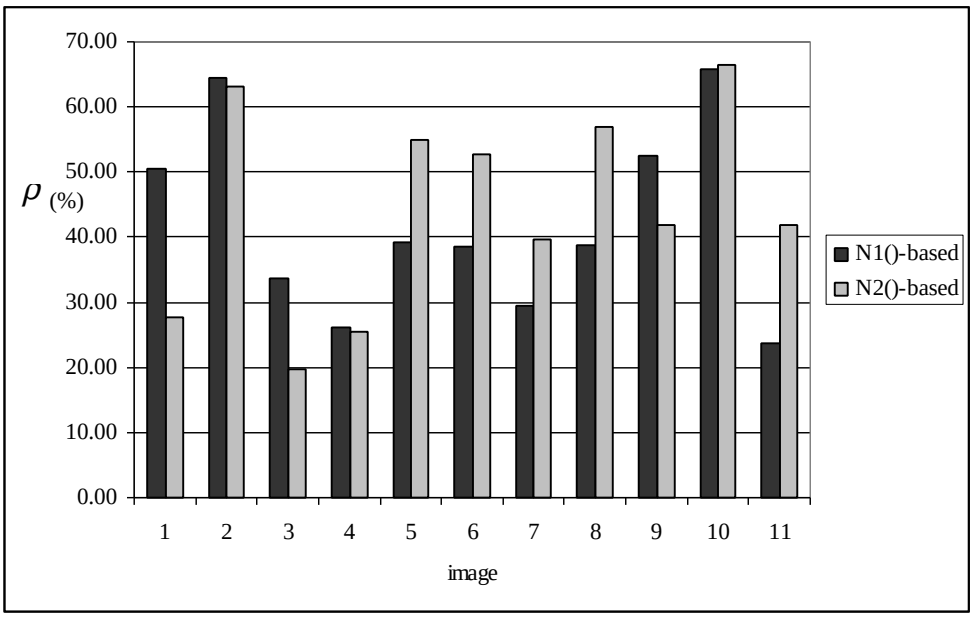


Figure 4.8: $\mathcal{N}_1(.)$-based model vs. $\mathcal{N}_2(.)$-based model: correlation coefficient ρ.

attention map derived from color images with two versions of the computational map of attention. The first version is computed from achromatic features like intensity and orientations, while the second version includes also the chromatic features. Note that both versions use the $\mathcal{N}_2(.)$ normalization strategy.

The basic idea behind this approach is that if the chromatic features do not influence the visual attention behavior, then the chromatic version of the model would have at most the same correlation with the human behavior than the achromatic version. Otherwise, the chromatic version of the model correlates with the human behavior at least as much as the achromatic version of the model does.

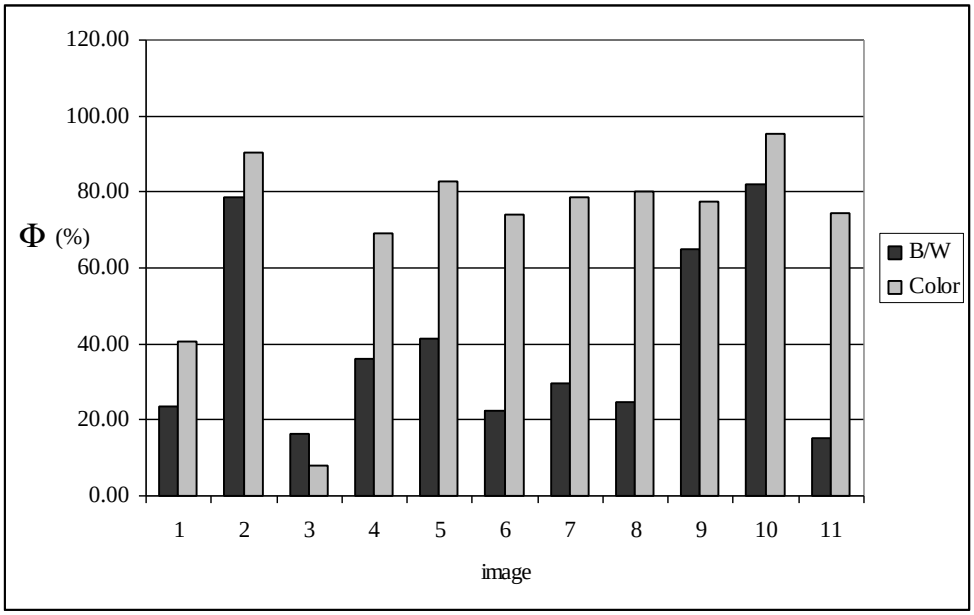


Figure 4.9: Color contribution to visual attention.

Figure 4.9 represents the comparison results between the achromatic and chromatic versions of the computational map of attention regarding their correlation with the human attention map, using the fixation-to-chance distance Φ. It can be observed that the chromatic-based attention map has better correlation with the human map than the achromatic-based map for almost the entire image set. Furthermore, the mean value of Φ passes from about 40% for the achromatic-based version to about 70% for the chromatic-based one, which represents an increase of about 75%.

4.3.4 Pop-out Effect

Image locations which differ from the rest of the image in a single feature dimension are considered to be pop-out stimuli. In visual search experiments, it has been shown that such stimuli are easily found by human subjects [149].

The computational model of visual attention is expected to have the same behavior regarding pop-out stimuli. If a location differs from the rest of the image according to a single feature, then the conspicuity map related to that feature is supposed to strongly highlight only this image location. Also, this conspicuity map is highly promoted by the normalization strategy used to integrate the different conspicuity maps into the final saliency map. Thus, the pop-out stimuli should give rise to high activities in the saliency map.

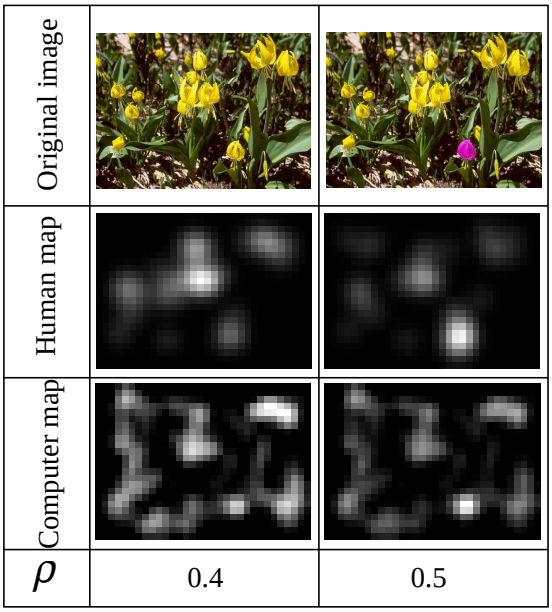


Figure 4.10: Color pop-out effect.

The human as well as the computational behavior on behalf of color pop-out stimuli are illustrated in Figure 4.10. In this experiment, we consider a scene which initially contains numerous colored locations (yellow flowers). For this initial image (first column in Figure 4.10), the human fixations are mostly concentrated around the image center, whereas the computational saliency is distributed over the entire image with the most active location at the top right corner of the image.

In a second experiment (second column in Figure 4.10), a manipulated version of the initial color image is used. The manipulation consists in coloring one yellow flower in red, without changing its luminance, which guarantees that changes in attention behavior is caused only by color changes. A clear shift of human

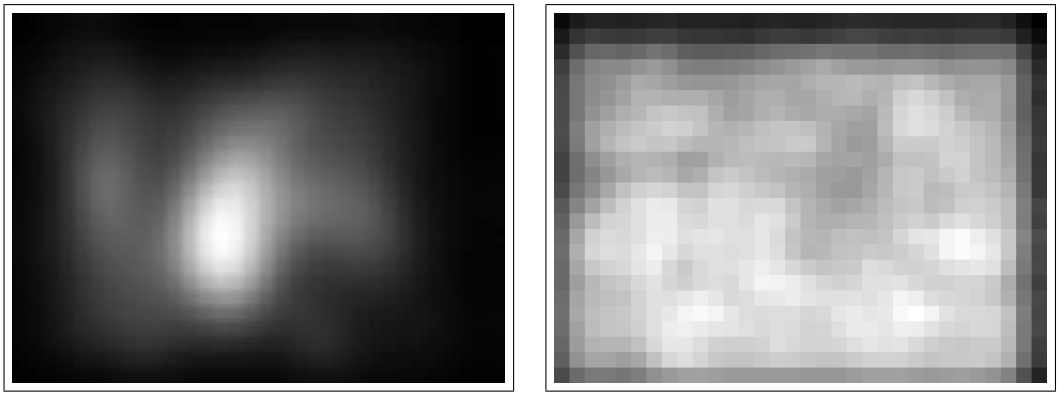
fixations towards the colored flower can be observed. The same colored flower constitutes the most salient image location for the computational model of visual attention. Also, the correlation coefficient of the human and computational maps of attention is clearly increased compared to the first experiment.

Note that we used the $\mathcal{N}_2(.)$ normalization strategy to compute the saliency maps. The corresponding parameters are set in the same way as in Section 4.3.2.

To conclude, the results of these experiments represent a further support to the hypothesis that states that the human attention mechanism has particular preferences to pop-out stimuli. They also demonstrate the existence of such behavior in the computational model of visual attention.

4.3.5 Image Center Effect

In the experiment configuration, there exist some human behaviors that influence the eye fixations, independently, of the presented stimuli. One of these behaviors is the image center bias [94], that is the tendency of humans to attend or to look at the central part of the presented image. To assess the existence of this effect, empirically, we carried out the following experiment.



(a) Mean human map (b) Mean saliency map

Figure 4.11: Image center effect. (a) represents the mean human map of attention computed from about 30 images and 20 human subjects. (b) depicts the mean computational saliency map computed from the 30 considered images.

The basic idea consists in analyzing the spatial distribution of the eye fixations recorded from several human subjects and several images. The presented images are of different nature (forest, traffic signs, fractal, ...). If the histogram of all recorded fixations exhibits a high concentration around the image center, that means that the center of images is often attended by the subjects. In our experiment we presented 30 different images to 20 human subjects. The histogram of all recorded fixation has, indeed, a remarkable concentration around

the image center, which confirms the high attendance of the center of the different images. To ensure that this high attendance of the center is not caused by the real existence of salient objects in that location of images, we compared the histogram of all recorded fixations with the mean saliency map, that is the mean image of all saliency maps computed from the same images. The relative uniform distribution of the mean saliency map guarantees the absence of specially salient objects in most of the considered images at that location. Both maps are given in Figure 4.11.

It can be concluded that the centered high concentration in the histogram of all recorded eye fixations is due to a bias of eye fixations towards the image center.

The Extension of the computational model of visual attention to consider the image center effect would certainly enhance the correlation between the human and the computational behavior regarding visual attention. However, for applications of the visual attention algorithm, the integration of the image center preference is not necessarily advantageous.

4.4 Chapter Summary

This chapter has dealt with the empirical validation of the saliency-based model of visual attention by comparing its performance with the human visual attention. Therefore, the validation method quantitatively evaluates the similarity between the human map of attention derived from eye fixation data with the computation attention map.

The experiments conducted with numerous human subjects and various color images have pointed to a high correlation between the human and the computation map of attention. The computed similarity scores were all positive, their mean value exceeded 40%, and reached 90% for some images.

In addition, the experiments have shown the superiority of the non-linear ($\mathcal{N}_2(.)$) combination of features into the final saliency map over the linear ($\mathcal{N}_1(.)$) combination method. Indeed, the linear-based computation map of attention correlates with the human map to about 40% in average, whereas the mean value of the similarity measure for the non-linear-based map reaches 70%.

The experiments have allowed also to quantitatively assess the contribution of color to visual attention. It has been observed that the consideration of chromatic features in the model of visual attention increases the similarity score of the computation and the human map of attention by 75%

Moreover, the preference of our computational model, like humans, to pop-out stimuli has been demonstrated by some targeted experiments.

Finally, the experiments have pointed out that the similarity between the human and computation map of attention could be increased further, if the image center bias was introduced to the visual attention model.

Chapter 5

Real-Time Visual Attention

5.1 Chapter Introduction

Implemented on a PC 900 MHz, a version of the saliency-based algorithm of visual attention, which considers color, intensity and depth runs at a frequency of only 2Hz. Indeed, due to its complexity, the real time operation of the reported model of visual attention needs higher computation resources than available in conventional processors. Until now, it was not achieved on a compact system.

Some previous works reported hardware models of visual attention implemented on fully analog VLSI chips [22, 67]. The authors considered, however, simplified versions of the saliency-based model of visual attention and implemented only small parts of it. In both works emphasis has been put only on the last stage of the attention model, namely, the winner-take-all (WTA) network. A complete real time software implementation of the saliency-based model of visual attention has been reported recently in [70]. However, the implementation has been carried out on a large system consisting of a 16-CPU cluster, named Beowulf. Involving 10 interconnected personal computers, the system might raise problems related to portability, power consumption, and price.

This chapter reports the first real time implementation of the complete saliency-based model of visual attention on a compact system, consisting of a low power, one board, highly parallel Single Instruction Multiple Data (SIMD) architecture, called ProtoEye [131, 118, 117].

5.1.1 Chapter Outline

The remainder of this chapter is organized as follows. Section 5.2 introduces the principles of SIMD architectures and stresses their suitability for low-level image processing. Section 5.3 describes the SIMD architecture ProtoEye which we used to implement a real-time attention system. The major issues of the implementation of the visual attention algorithm on ProtoEye are presented in

Section 5.4. Section 5.5 reports experimental results which validate the different steps of the implemented visual attention system. Section 5.6 quantitatively assesses the performance of the system by deeply analyzing the different operations needed for computing visual attention and gives some perspectives on how to further improve the architecture to cope with higher frequencies, more scene features and larger images. Finally, the chapter is summarized in Section 5.7 and some conclusions are stated.

5.2 Vision Chips

The need for real-time sensory information processing is increasing in the field of computer vision. Vision-based robot navigation, autonomous mobile vehicles, high speed quality inspection are typical applications which require high speed visual feedback. Therefore, various highly parallel vision architectures have been developed. They can be roughly classified into three major categories: fully analog; fully digital; and mixed analog-digital architectures.

5.2.1 Fully Analog Vision Chips

This category of vision chips refers to those vision systems that use only analog circuits to realize image processing devices. Most of these vision systems integrate the photo-detecting elements and the processing elements on the same chip, bringing to reality the concept of "Smart Sensors". Generally, a processing element is available for each pixel, which leads to a parallel processing of the entire image. The CSEM motion detector chip for pointer devices [8] and the orientation detector chips of Standley [139] are examples of these fully analog vision chips. For a more complete survey of this category of vision chips, the reader is referred to [98].

Compared to conventional vision systems which are composed of a camera and a personal computer, the described analog vision chips have the following advantages.

- **Speed**: Due to the parallel processing capability of the vision chips, the computation time of image processing algorithms is drastically reduced when using such vision chips.
- **Size**: Very compact vision systems can be realized, when using single chip implementation of image processing algorithms.
- **Power dissipation**: Vision chips often use analog circuits which operate in sub-threshold region, yielding low power consumption.

The analog vision chips are, however, at a disadvantage regarding the following points.

- **Reliability of processing**: The precision in analog VLSI systems is affected by many factors which are not well controllable. Thus, if the algorithm does not account for these inaccuracies, the processing reliability may be severely affected.

- **Resolution**: The processing element circuits occupy a large portion of the pixel area. Therefore, vision chips have a low fill-factor and thus a low resolution.

- **Programming**: The fully analog vision chips are not general purpose, since they are not programmable to perform different vision tasks. This inflexibility is particularly undesired during the development of vision systems.

5.2.2 General Purpose Digital SIMD Architectures

Faced with the rigidity of the fully analog vision systems, numerous research groups worked on developing more flexible vision chips, which gave birth to some general purpose digital vision chips [14, 120, 81]. These vision chips are built around Single Instruction Multiple Data (SIMD) architectures, that is an array of identical processing elements, each executing the same instruction on one element of an array of data. Typical constituents of the processing elements of such chips are Arithmetic Logic Unit (ALU), local memory and input/output interfaces. Neighboring processing elements can also communicate in order to permit the implementation of neighborhood-based operations (like edge detection). Figure 5.1 depicts the architecture of a 64×64 fully digital vision chip reported in [81].

Though, they completely resolve the programmability problem, the fully digital vision chips have a serious drawback related to image filtering using large kernels. In fact, digital SIMD architectures are not suitable for operations which require the consideration of a large neighborhood, since each processing element is connected only to its direct neighbors.

5.2.3 Mixed Analog-Digital SIMD Architectures

Mixed analog-digital SIMD architectures have been conceived to overcome the weaknesses of both fully analog and fully digital vision chips. On one hand, mixed analog-digital SIMD architectures are fully programmable, since they have the same digital constituents as the fully digital vision chips. On the other hand, the analog part of the mixed architecture properly solves the problem of image filtering using large kernels. Some mixed SIMD architectures contain a diffusion network that implements lowpass as well as highpass spatial filtering capabilities with a negligible complexity, regardless of the size of the filter.

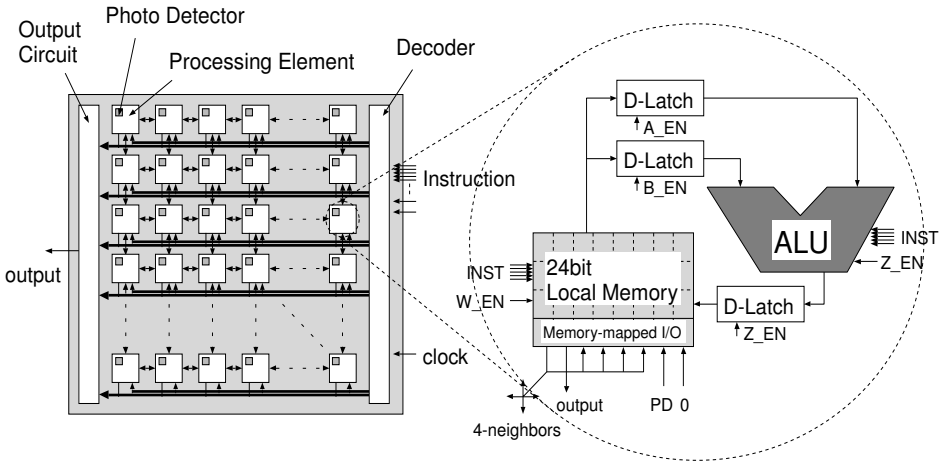


Figure 5.1: General purpose digital vision chip (from [81]).

Given the advantages of such vision chips, a mixed analog-digital SIMD architecture, called ProtoEye [131] and provided by CSEM [36] has been used to realize a real-time visual attention system.

5.3 ProtoEye: SIMD Machine for Image Processing

5.3.1 Overview of the Architecture

The complete vision system is composed of a CMOS imager (352×288 pixel), a video output, a general purpose microcontroller, and 4 ProtoEye chips (Figure 5.2). A 64×64 pixel subimage is grabbed from the image provided by the camera, and transferred to the ProtoEye architecture by means of a Dynamic Memory Access (DMA) interface. The same DMA interface is used to transfer the final results from ProtoEye to the external memory, which is interfaced to the video output. It is noteworthy that the data transfer between the DMA and the ProtoEye (and vice versa) is achieved in a serial manner.

The ProtoEye architecture is a SIMD machine composed of a microcontroller and a 64×64 array of identical processing elements (PE). The microcontroller consists of a 4 MHz clocked RISC processor implemented on an FPGA.
A PE, which is associated to a single pixel and is connected to its 8 neighbors, is composed of a digital part and an analog one (Figure 5.3).

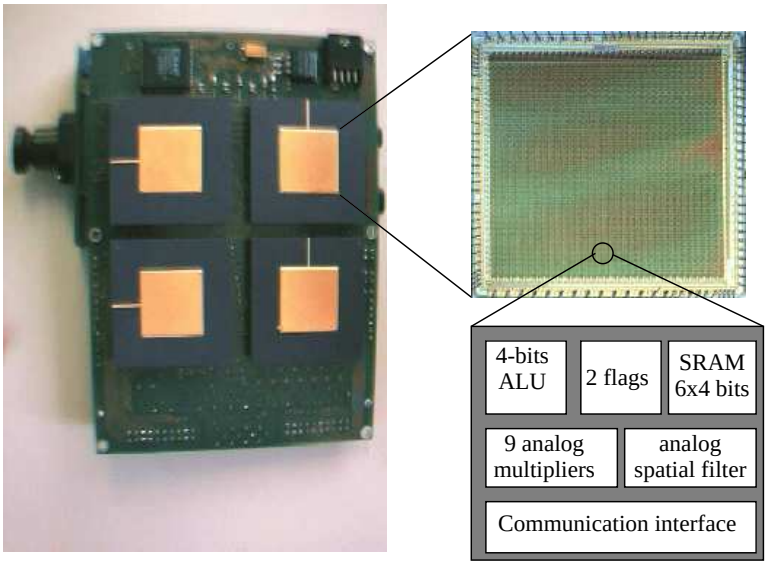


Figure 5.2: ProtoEye platform. This vision platform is composed essentially of four ProtoEye chips, a CMOS camera, a microcontroller and a video output.

5.3.2 The Digital Part of ProtoEye

The digital part of a PE is organized around an internal 4-bit D-bus (D[3:0]). It contains a 4-bit ALU, which has as input the D-bus and the accumulator. The ALU instruction set includes all logical functions, addition, subtraction, shifts of the accumulator content and comparison. The flag $F1$ masks conditional operations. The six 4-bits registers are used to keep temporary results within the processing element. In digital mode, transfers between neighboring PEs can be performed by shifting the accumulator content. Despite reduced resources, this part of ProtoEye offers a high programming flexibility.

5.3.3 The Analog Part of ProtoEye

The analog part of each PE (shaded area on Figure 5.3) is connected to the digital part through Analog to Digital (A/D) and Digital to Analog (D/A) converters. Its essential component is the analog spatial filter, which is based on a diffusion network, made of pseudo-conductances connecting the PEs [156]. The Diffusion network represents a powerful tool for image filtering since the complexity of such operation is O(1), whatever the size of the filter. It implements an exponential spatial filter H as defined in Equation 5.1 (for the 1D case).

$$H(i) = k \cdot e^{-\frac{i}{\lambda}} \tag{5.1}$$

The input of the spatial filter is the content of the register RAM5, converted by the D/A converter. Its output is a lowpass filtered version of the input image,

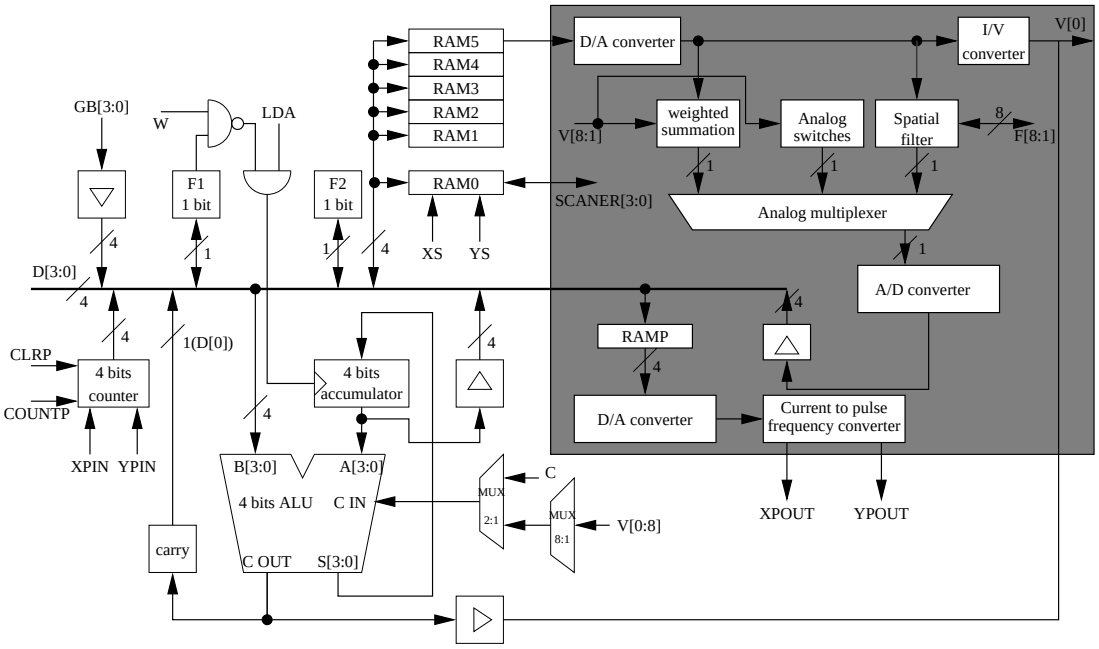


Figure 5.3: ProtoEye: Architecture of a single processing element (PE). A PE is composed of a digital part (non shaded area) and an analog part (shaded area). The digital part consists of a 4 bit ALU, an accumulator, six registers (RAM0 ... RAM5) and two 1 bit flags (F1 and F2). The main component of the analog part is the analog spatial filter which is based on a diffusion network.

which cut-off frequency is controlled by two external voltages V_R and V_g. Taking the difference between original image and its filtered version results in a highpass filtered version of the input image.

Note that the 64×64 array of PEs is implemented on four chips.

Characterization of the Analog Filter

As mentioned above the diffusion length of the analog filter is controlled by two external voltages V_R and V_g. In fact, the user programs two 12 bits registers whose content are then converted to analog currents by an D/A converter. The obtained currents determine the diffusion length λ of the filter. This section aims at, empirically, finding out the relation between the digital values entered by the user and the behavior of the analog filter of ProtoEye.

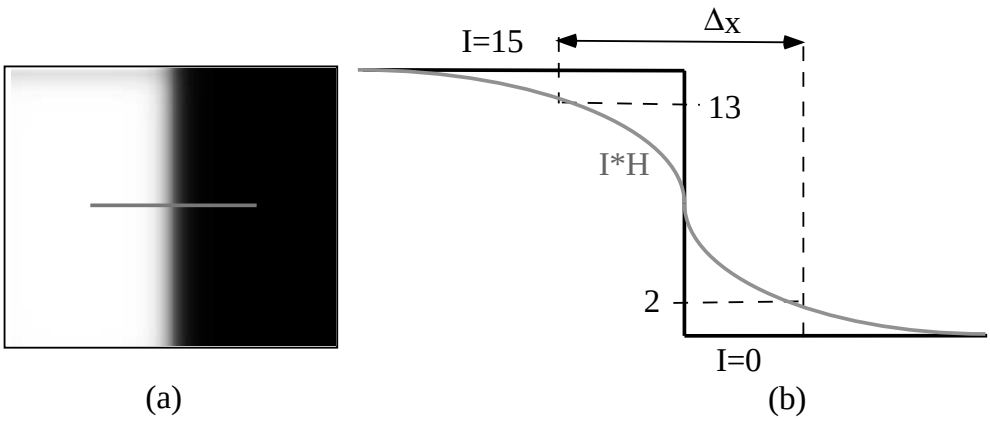


Figure 5.4: Characterization of the analog filter through analyzing horizontal profiles of filtered step images. (a) smoothed step image. (b) profile analysis.

Therefore, the following experiment has been undertaken. First, a step image is acquired and then binarized (i.e. 0 and 15) by ProtoEye. Second, the numerical values V_R and V_g of the A/D converter inputs are varied and for each value, the corresponding filtered image is stored. Finally, one can determine the parameter λ of the simulated exponential filter through analyzing horizontal profiles of the filtered step images.

Figure 5.4 illustrates the described experiment. In this illustration, we determined, on an horizontal profile of the filtered step image, all pixels whose values lie in the interval [2..13]. Let Δx be the length (in pixel) of this portion of the profile. Since the spatial filter has an exponential form, the relation between λ and Δx can be formulated in accordance with Equation 5.2.

$$\begin{aligned} k \cdot e^{\frac{-\Delta x/2}{\lambda}} &= 2 \qquad \Leftrightarrow \\ \lambda &= -\frac{\Delta x}{2 \cdot \ln(\frac{2}{k})} \qquad \text{(where } k = 7.5\text{)} \end{aligned} \tag{5.2}$$

The experiments have shown that the diffusion length λ is quasi proportional to the difference $(V_G - V_R)$. Figure 5.5 depicts the variation of λ with V_G while keeping V_R constant ($V_R = 700$). Note that V_R and V_G are scalar values with no unit.

These information about the diffusion network are essential for the realization of the real-time attention system on ProtoEye, since spatial filtering is a fundamental issue in our algorithm.

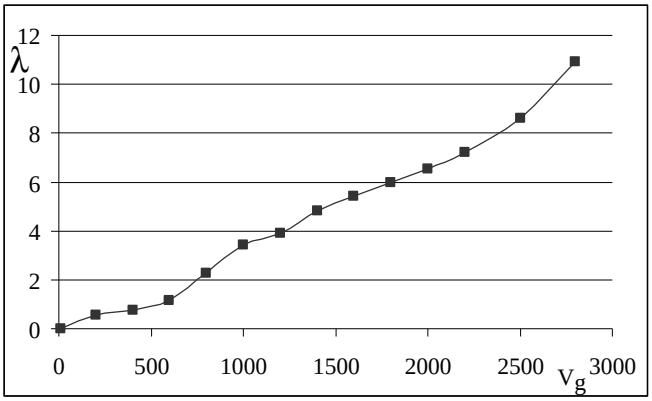


Figure 5.5: Variation of the diffusion length λ (in pixel) with V_G while V_R is kept constant ($V_R = 700$).

5.4 Implementation Issues

This section reports the implementation of the saliency-based model of visual attention reported in Chapter 2.4 on the described SIMD architecture ProtoEye. We report a version of our implementation of visual attention on ProtoEye that considers two image features, namely intensity and motion. In a first step a conspicuity map related to intensity is computed. Then, we compute a map related to scene motion. Finally, the two maps are integrated into a final saliency map and the most salient locations are derived.

5.4.1 Intensity Conspicuity Map

Let us recall that the computation of a conspicuity map from a given feature map relies on two major operations, the center-surround that computes the interme-

diate multiscale conspicuity maps, and the integration process which combines these multiscale maps into a feature-related conspicuity map.

Center-Surround Transformation

We implement this transformation by means of a multiscale difference-of-exponential filter. Practically, a nine level exponential pyramid $\mathcal{P}$ is built by progressively lowpass filtering the intensity image I using an exponential filter H. Formally, this pyramid is defined according to Equation 5.3.

$$\begin{aligned} \mathcal{P}(0) &= I \\ \mathcal{P}(i) &= \mathcal{P}(i-1) * H \end{aligned} \tag{5.3}$$

where $(*)$ refers to the spatial 2-dimensional convolution operator.

It is noteworthy that the levels of the pyramid $\mathcal{P}(i)$ have the same resolution. Indeed, the multi-resolution approach efficiently realizes the computationally complex concept of multiscale at cost of spatial accuracy [87]. Since the complexity issue is solved thanks to the diffusion network on one hand and to the parallel nature of ProtoEye on the other hand, multi-resolution looses, in this case, its raison d'être.

Six multiscale conspicuity maps $C_{1..6}$ are then computed from the exponential pyramid as absolute differences between fine scales (center) and coarse scales (surround), according to Equation 5.4.

$$\begin{aligned} C_1 &= |\mathcal{P}(2) - \mathcal{P}(5)|, \quad C_2 = |\mathcal{P}(2) - \mathcal{P}(6)| \\ C_3 &= |\mathcal{P}(3) - \mathcal{P}(6)|, \quad C_4 = |\mathcal{P}(3) - \mathcal{P}(7)| \\ C_5 &= |\mathcal{P}(4) - \mathcal{P}(7)|, \quad C_6 = |\mathcal{P}(4) - \mathcal{P}(8)| \end{aligned} \tag{5.4}$$

These multiscale conspicuity maps are sensitive to different spatial frequencies. Fine maps (e.g. C_1) detect high frequencies and thus small image regions, whereas coarse maps, such as C_6, detect low frequencies and thus large objects.

The difference-of-exponential filter ($\mathcal{D}o\mathcal{E}xp$) is more suitable for ProtoEye than the difference-of-gaussians one ($\mathcal{D}o\mathcal{G}$) used in the original model. The analog diffusion network, which implements exponential filtering, efficiently computes the exponential pyramid in a negligible time. Also, this modification does not affect the results of the conspicuity transformation, compared with the original model, since both filters ($\mathcal{D}o\mathcal{E}xp$ and $\mathcal{D}o\mathcal{G}$) are very similar [41], as shown in Figure 5.6. The similarity of both filters guarantees the fidelity of the modified conspicuity operator to the original one. For comparison purposes, the relation between the diffusion length λ of an exponential filter and a standard deviation σ of a gaussian one is given by Equation 5.5 ([41]).

$$\frac{\sigma}{\lambda} = \frac{5}{2\sqrt{\pi}} \tag{5.5}$$

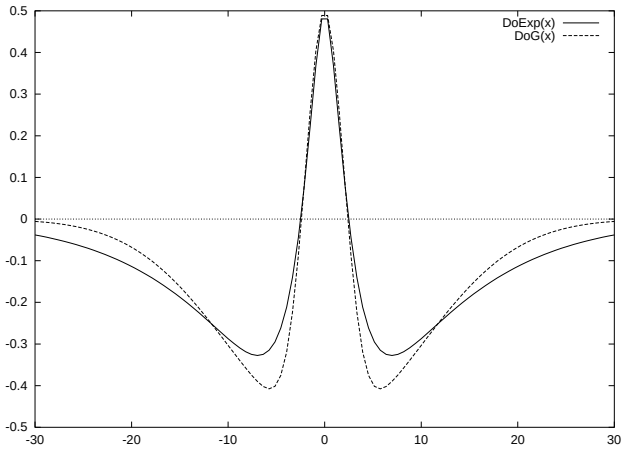


Figure 5.6: $\mathcal{D}o\mathcal{E}xp$ versus $\mathcal{D}o\mathcal{G}$.

The number (six) and the size (4 bits) of the internal registers of ProtoEye represent an additional constraint to deal with in our implementation. Six registers are not enough to store the nine levels of the exponential pyramid. To overcome this constraint, we keep only two levels of the pyramid, at once, in the registers: a center level (level 2, 3 and 4 successively) in RAM2 and a surround level (level 5, 6, 7 and 8 successively) in RAM3. Once a conspicuity map is computed, the corresponding pyramid levels are replaced by the two levels needed to compute the next conspicuity map (see Figure 5.7).

Note that the first center level ($\mathcal{P}(2)$) and the first surround one ($\mathcal{P}(5)$) of the exponential pyramid are computed directly using two exponential filters with different diffusion lengthes H_2 and H_5 as defined in Equation 5.6.

$$\begin{aligned} H_1 &= H \\ H_i &= H_{i-1} * H \end{aligned} \tag{5.6}$$

Due to the same constraint, each computed conspicuity map is transferred to external memory (*toMem()* on Figure 5.7), after the normalization (*Norm()* on Figure 5.7). The normalization of the different maps is described below.

The Integration of the Multiscale Maps

The six multiscale conspicuity maps have to be combined, in a competitive way, into a unique feature-related conspicuity map. Taking advantage from the analog diffusion network of ProtoEye, we implemented the normalization strategy $\mathcal{N}_2(.)$ - iterative non-linear normalization - presented in Section 2.4.5. Let us recall that this strategy relies on simulating local competition between neighboring conspic-

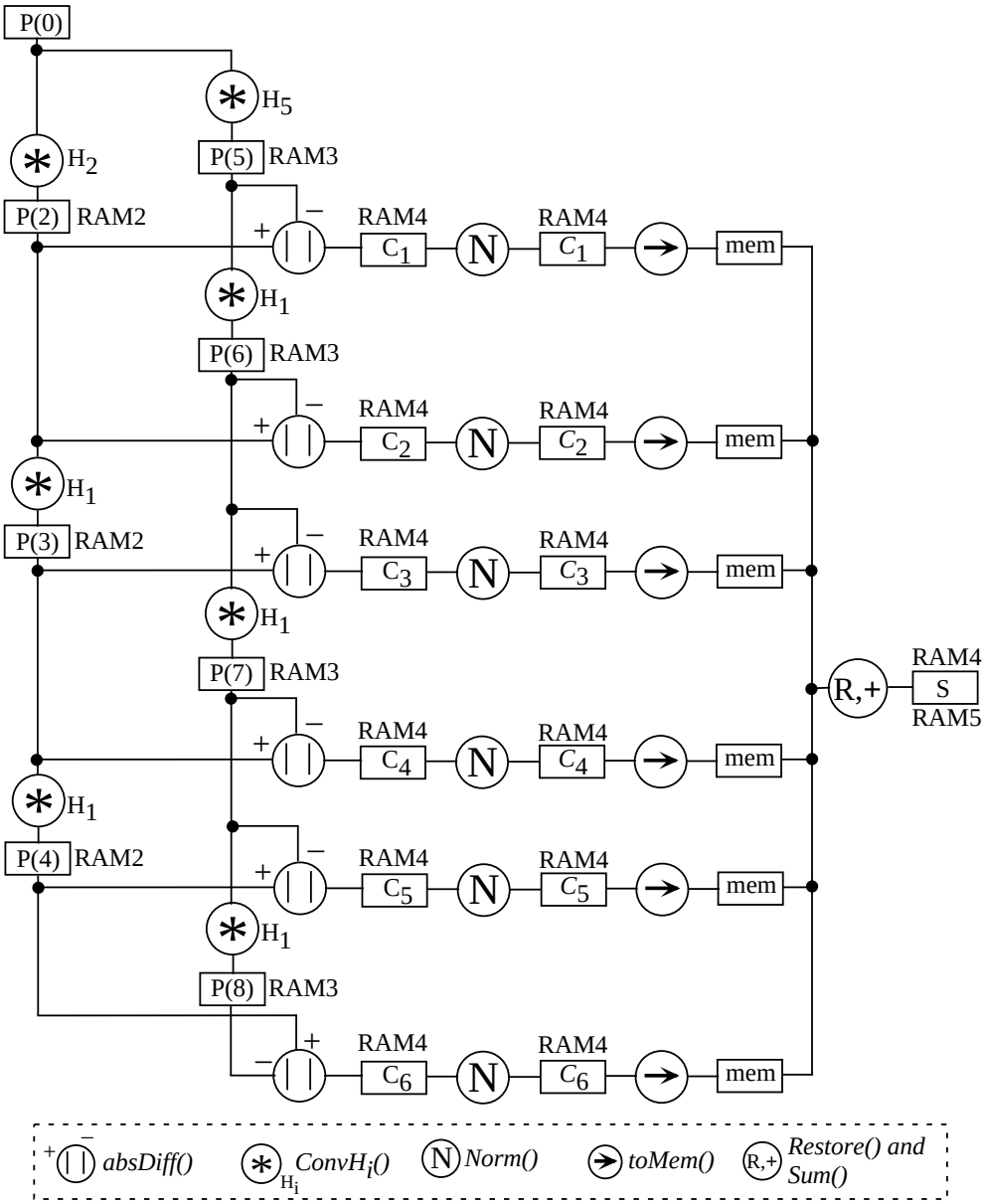


Figure 5.7: Resources allocation while computing the visual attention model for a single scene feature (intensity). From a grey-level image, an exponential pyramid is built by progressively lowpass filtering the input image, using the function $ConvH_i()$ (as defined by Equation 5.6). The absolute difference between different levels of the pyramid gives the six multiscale conspicuity maps C_1, C_2, C_3, C_4, C_5, and C_6, which are successively normalized and stored into the external memory. These maps are then restored to internal registers (RAM4 and RAM5) and a double precision summation of all maps is computed. This sum, divided by 8, corresponds to the saliency map.

uous locations. Spatially grouped locations which have similar conspicuities are suppressed, whereas spatially isolated conspicuous locations are promoted.

First, each map is scaled to values between 0 and 15 in order to remove modality-dependent amplitude differences. Each map is then iteratively convolved by a large 2D difference-of-exponentials filter $\mathcal{D}o\mathcal{E}xp$ (the original version of the normalization strategy uses a $\mathcal{D}o\mathcal{G}$ filter). The negative results are clamped to zero after each iteration, which guarantees the nonlinearity of the normalization method (more details about the parameters of the normalization are given in Section 5.5).

At each iteration of the normalization process, a given intermediate conspicuity map C is transformed according to Equation 5.7.

$$C \leftarrow |C * \mathcal{D}o\mathcal{E}xp|_{\geq 0} \tag{5.7}$$

where $(*)$ is the convolution operator and $|.|_{\geq 0}$ discards negative values. Note that the normalization operator is applied before the maps are transferred to the external memory.

The final conspicuity map $\mathcal{C}_I$ related to intensity is then computed in accordance with Equation 5.8.

$$\mathcal{C}_I = \frac{C_1 + C_2 + C_3 + C_4 + C_5 + C_6}{8} \tag{5.8}$$

Practically, the normalized maps stored in the external memory are restored into two internal registers (RAM4 and RAM5) for the final merge which mainly consists of a double precision summation and a division by 8.

5.4.2 Motion Conspicuity Map

We propose a simple method that permits the detection of changes in a scene. We suppose that the camera is static and changes in the scene are due to objects motion. A simple difference between two successive frames is sufficient to highlight the moving objects, under these assumptions.

Practically, the previous image is kept in the internal memory of ProtoEye and once a new image is acquired, an absolute difference between the two images is computed, giving rise to the motion-related conspicuity map $\mathcal{C}_M$. The new image then replaces the previous one in the internal memory.

5.4.3 Saliency Map

Given the intensity-related conspicuity map $\mathcal{C}_I$ and the motion-related one $\mathcal{C}_M$, the final saliency map $\mathcal{S}$ is computed in accordance with Equation 5.9.

$$\mathcal{S} = \frac{\mathcal{C}_I + \mathcal{C}_M}{2} \tag{5.9}$$

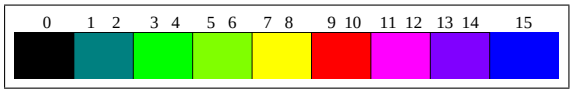
5.4.4 Detection of the Spots of Attention

The final step of the task consists in selecting the most salient parts in the image. Therefore, we suggest a k-Winner-Take-All (kWTA) network based on a large difference-of-exponential filter. The kWTA is iteratively applied on the saliency map. It separates the image locations into two categories, winners and losers, depending on their saliency activities. One or several spots are detected, depending on the diffusion length of the used $\mathcal{D}o\mathcal{E}xp$.

So far, the final result of the developed visual attention system is a binary image which displays the most salient image locations. This image is transferred to the external memory, which permits its use in subsequent tasks.

5.5 Experimental Results

In this section we report experiments that assess the proposed implementation of the different steps of the visual attention model discussed in Section 5.4.



Pseudo-coloring of the 4-bit gray level (0 .. 15)

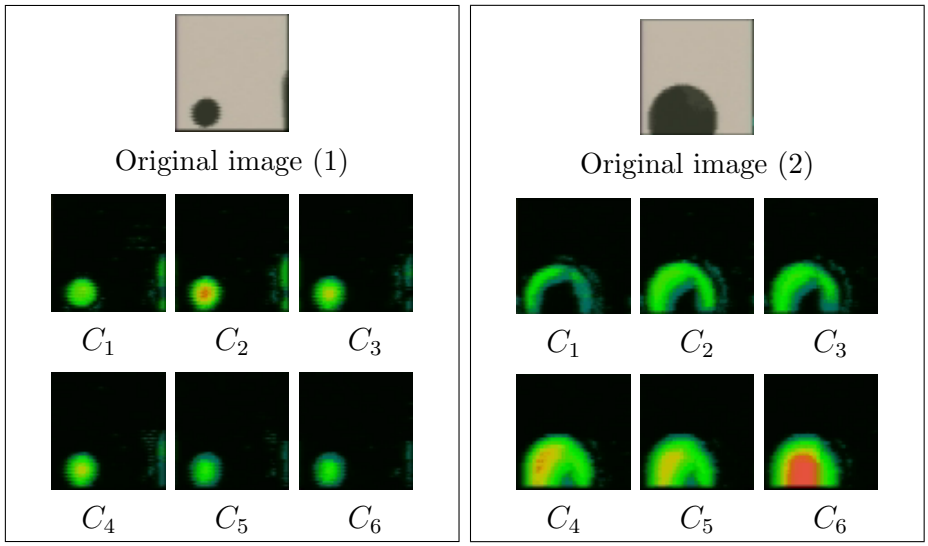
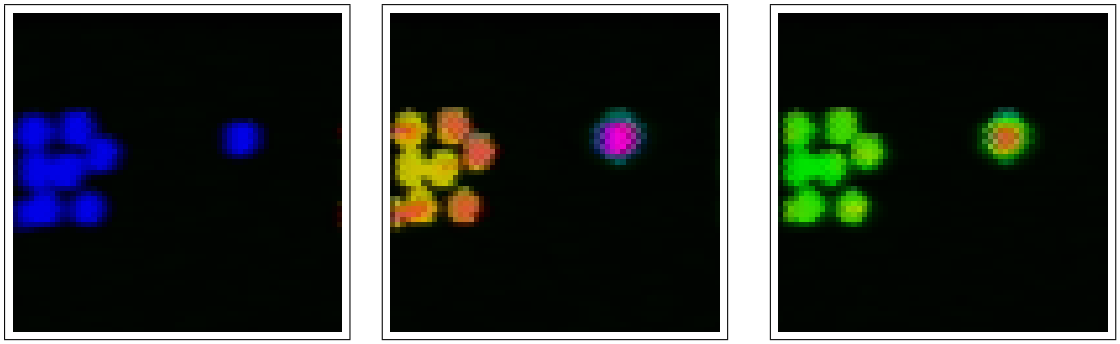


Figure 5.8: Multiscale conspicuity transformation.

The first experiment (Figure 5.8) refers to the operation of the multiscale channel. Two different scene images have been considered. For each image, the six multiscale conspicuity maps (C_1 .. C_6) are computed. Note that the exponential filter H used to build the pyramid has a diffusion length $\lambda = 1.5$ for both images. The activity of the conspicuity maps is pseudo-colored according

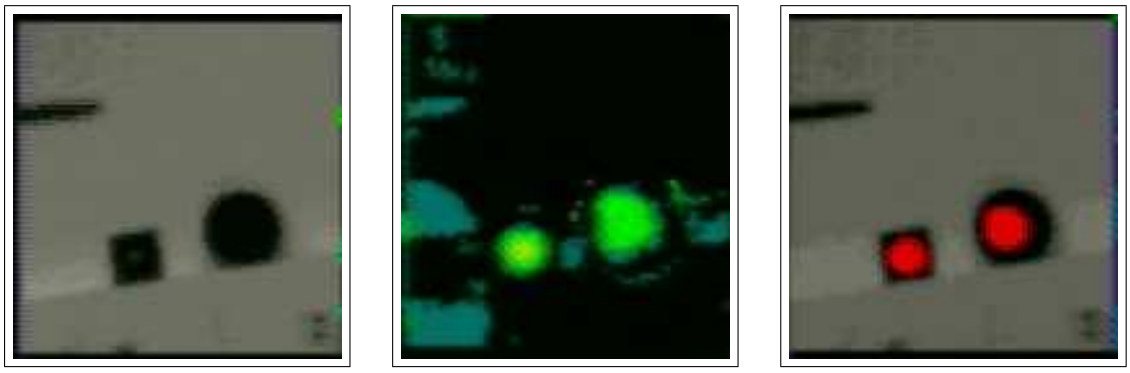
to the color palette of the same figure (top). The first image (left) consists of a small black disc on a white background. The conspicuity map C_2 has the highest response among the six maps. Due to the larger size of the disc on the second image (right), C_6 is the conspicuity map that contains the highest activity.

To summarize, this experiment validates the implemented multiscale conspicuity transformation, since the different conspicuity maps are sensitive to different spatial frequencies: fine maps detect high frequencies and thus small image regions, whereas coarse maps detect low frequencies and thus large objects.



(a) Conspicuity map (b) 1. iteration (c) 2. iteration

Figure 5.9: Iterative normalization of conspicuity maps.



(a) Original image (b) Saliency map (c) Spots of attention

Figure 5.10: Detecting the most salient locations in a grey-level image. First, we compute a saliency map from the intensity feature. Then, we apply a maximum network on the saliency map in order to detect the most salient locations or spots of attention. These spots are marked up with red stains.

The second experiment (Figure 5.9) refers to the iterative normalization process. A conspicuity map is considered, which contains on one hand a set of

spatially grouped spots and on the other hand a spatially isolated spot. We then iteratively convolve this map with a $\mathcal{D}o\mathcal{E}xp$ filter whose diffusion lengths λ_{On} and λ_{Off} are set to 1.5 and 6.5, respectively. The activity of the maps are pseudo-colored using the color palette on Figure 5.8 (top). The spatially grouped activities are progressively suppressed compared to the isolated spot. This clearly shows the competition between neighboring conspicuous locations and thus validates the implemented normalization method $\mathcal{N}_2(.)$.

The third experiment (Figure 5.10) refers to the last stage of the attention model, namely, the kWTA network. From a gray level real image (left), a saliency map (middle) is first computed. The kWTA is then applied on it, using the following parameter settings: $\lambda_{On} = 1.5$; $\lambda_{Off} = 6.5$; and $p = 3$ (p is the number of iterations). The resulting spots (winners) are colored in red and are mapped onto the original image (right).

The final experiment (Figure 5.11) depicts the conspicuity maps related to motion and intensity, the corresponding final saliency maps as well as the spots of attention computed from an image sequence. In the absence of motion in the scene, the intensity contrast determines the saliency of locations, whereas the appearance of moving objects in the same scene considerably changes the constellation of the detected spots of attention. This example clearly shows the influence of the motion cue on the final saliency map and consequently on the detection of the spots of attention.

To conclude, the experiments clearly validate the effective operation of the saliency-based model of visual attention implemented on ProtoEye.

5.6 Performance Analysis and Perspectives

5.6.1 Performance Analysis

It is obvious that the computation of the intensity-related conspicuity map is the most time consuming step in the two-features model of visual attention. The computation of the motion map is instantaneous, since it consists only in a single arithmetic operation.

Thus, to analyze the performance of the system, we focus on the computation of the sole intensity-related conspicuity map. Therefore, we measure the computation time of this process as well as the computation time of each single step within the task of computing the intensity conspicuity map. These measurements are represented in Table 5.1.

The complete computation of an intensity conspicuity map takes 71.1 ms, leading to a computation frequency of 14 images/second.

Let us consider now the individual contribution of each function to the complete computation time. It can be observed that the image grab constitutes over 43% of the entire computation time.

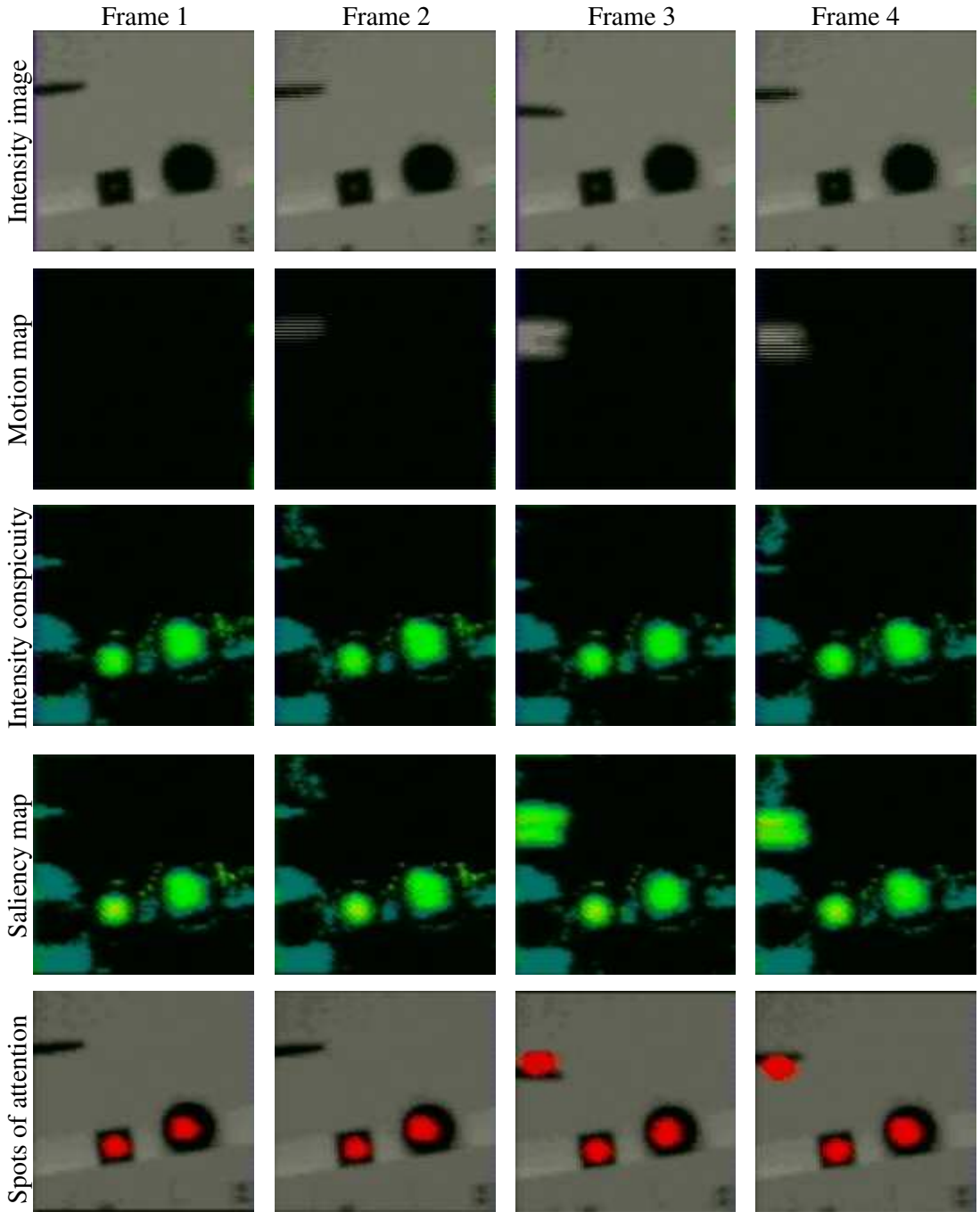


Figure 5.11: Spots of attention from intensity and motion.

Functions	Execution time of operations (ms)					Total (ms)
	Conv()	Arithm*	Norm()	toMem()	restore()	
GrabImage()						30.75
$C_1, C_2, .., C_6$	4.5	0.024	12.72	10.68	10.68	38.6
Final map		0.015		1.78		1.79
All						71.1

* absDiff() and Sum()

Table 5.1: Computation time of the complete process and of the single operations. *Conv()* is the smoothing function, which is used to compute the exponential pyramid $\mathcal{P}(i)$. *absDiff()* and *Sum()* are the two arithmetic operations needed to compute the six conspicuity maps and the final map respectively. *Norm()* is the normalization function. *toMem()* and *restore()* are responsible for the image transfers between the processing elements and the external memory and vice versa.

The computation of the final conspicuity map, which consists in a double precision summation of the multiscale maps and a division, represents less than 3% of the processing time. The time required for the computation of the six multiscale conspicuity maps exceeds 54% of the entire computation time.

Due to its large contribution to the computation time of the entire process, the computation time of the multiscale maps is closely analyzed. Figure 5.12 illustrates the distribution of the computation time of a single multiscale conspicuity map over the required operations. It is noteworthy that the operations, which require an image transfer between the processing elements and the external memory (*Norm(), toMem(), restore()*), hold the major part of the computation time because, as mentioned above, the data transfer from/to the processing elements is achieved in a serial manner.

Norm() requires this transfer in order to compute the global maximum of the maps, which is used in their normalization. The two operations *toMem()* and *restore()* require over 55% of the computation time of a conspicuity map. As mentioned above, these two functions allow the storage of individual conspicuity maps in external memory for lack of enough internal registers.

Note that one additional internal register (let us call it RAM6) would be enough to avoid these two image transfer based functions. It would allow to reduce the computation time of a conspicuity map by 55% and of the entire computation time by 30%. This computation time reduction would yield an operation frequency of 19 images/second.

To summarize, the computation time of the entire process is distributed as follows:

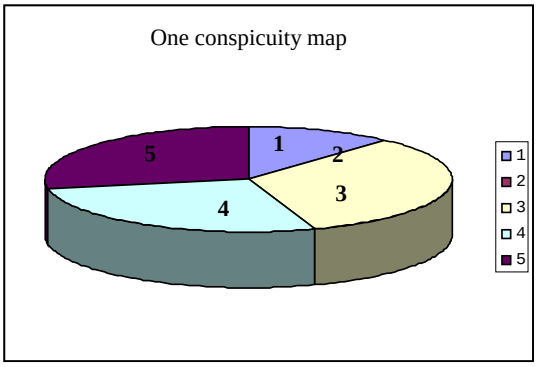


Figure 5.12: Distribution of the computation time over the operations which are used to compute a single multiscale conspicuity map. 1) Conv(), 2) absDiff(), 3) Norm(), 4) toMem(), and 5) restore().

- The image grab and its transfer to ProtoEye represents 43% of the cycle time.
- Of the remaining cycle time used for computation, the image transfer needed to store and restore temporary results holds 80% of the time.
- Of the same remaining cycle time used for computation, real processing which takes place in ProtoEye represents less than 20% of the time.

5.6.2 Perspectives

We discuss now how the presented architecture can be further improved to cope with a higher image rate, more scene features and larger images. Considering above performance analysis, it appears that the image grab and transfer time makes more than 40% of the cycle. This part is composed of a large delay for image acquisition and a single transfer cycle. The first can be eliminated by image buffering and the second can be counted with the other many image transfers required for the conspicuity computation. Under this hypothesis, the processing performance reaches a frequency of about $f_0 = 20$ Hz.

We consider then the extension from 1 to n features and a possible speed-up of the architecture operating frequency by a factor of x. The realistic hypothesis that the system scales according to a linear rule with x and inverse rule with n leads to the following formula for the new image rate f:

$$f = f_0 \frac{x}{n} \tag{5.10}$$

Given that the current circuits are running at a frequency of 4 MHz and that it is not unrealistic to consider an operation at 40 MHz, we count on a possible

speed-up of $x = 10$ giving enough room for increasing n by addition of new features or/and increasing f by operating the system at a higher frequency.

Considering now the system extension towards larger images, it is useful to consider from above that each cycle consists approximately for 20% in a parallel processing part and for 80% in a part devoted to a data transfer which has serial character. The duration of the first part is thus independent of the image size while the second scales according to a quadratic rule with a one-dimensional image scale s. It appears therefore that the system activity becomes increasingly dominated by data transfers when the image size is increased. A means to overcome this obstacle is the introduction of some fast and parallel data transfer procedure, like for example the simultaneous access to several image lines. Assuming a parallelism of degree P, the formula for the new cycle time T of a s scaled system with respect to the original cycle time $T_0 = 50$ ms is given by Equation 5.11.

$$T = T_0(0.2 + 0.8\frac{s^2}{P}). \tag{5.11}$$

In order to maintain the original performance of above described system with a larger image of 256×256 ($s = 4$) for example, it appears that a fast data transfer is required and the formula tells that a degree of parallelism of $P = s^2 = 16$ is required at least.

It appears finally that the presented architecture can cope with higher image rates and more scene features, mainly by boosting the speed of the processor. Coping with larger image sizes requires the rather simple replication of the elementary processing elements, but in addition, a parallel data transfer mechanism of degree P must be developed.

5.7 Chapter Summary

This chapter has reported the first real-time implementation of the saliency-based model of visual attention on a compact system consisting of a low power, one board, highly parallel SIMD architecture. Conceived for general purpose low-level image processing, the fully-programmable SIMD machine consists of an array of mixed digital-analog processing elements that offer high-performance functionalities for implementing the various functions appearing in the model of visual attention. The current prototype processes 64×64 images at a rate of 14 images/second, which allows the use of visual attention in real-time applications related to computer vision. Extensive performance analysis has confirmed the strengths of the architecture, but also has shown its high performance potential. Future designs, where a 10 times increase in computing performance seems straightforward, allow the integration of further image features into the attention model and faster image rates. Due to the parallel architecture, the extension to

larger images is also possible at the cost of simple hardware replication, provided that the serial image transfer is upgraded into a parallel form.

Chapter 6

Application of Visual Attention to Computer Vision

6.1 Chapter Introduction

Vision is considered as an attentive process [151] which is achieved in two phases: a parallel pre-attentive phase and an attentive one. The first phase aims at computing the saliency map, whereas the second phase consists in sequentially analyzing in more details the most salient locations of a scene. So far, the saliency-based model of visual attention implements the first phase of the vision process.

Since visual attention selects a reduced set of potentially informative locations of a scene, numerous computer vision applications might benefit from this mechanism. Indeed, higher level vision tasks, which basically have high computational complexity, can be speeded up by allocating the available computational resources for only a small part of the visual information. Furthermore, the saliency-based model of visual attention provides, through its multi-modal and hierarchical structure, relevant information about the salient parts of the scene. This information can be used in adapting the behavior of vision tasks according to the characteristics of single scene parts, which might increase the performance of these tasks.

The current chapter reports our attempts to implement some aspects of the second phase of the vision process - the attentive phase - by using the visual attention mechanism to guide some computer vision tasks.

6.1.1 Chapter Outline

The remainder of this chapter is organized as follows. Section 6.2 reports an attention-based color image compression method that takes advantage from the saliency-based model of visual attention to adaptively determine the number of

bits to be allocated for coding image regions according to their salience. Then, a color image segmentation method which extensively relies on the information provided by the visual attention algorithm, like salient regions, salient scales, and salient features, is described in Section 6.3. Section 6.4 discusses the usefulness of pre-attentive scene information for solving the problem of object tracking in a dynamic scene and gives some ideas how the proposed attention-based tracking method can be used in the field of visual robot navigation. The last computer vision application considered in this thesis is related to driver assistance. Indeed, Section 6.5 reports an attention-based method for traffic sign detection and recognition. Finally the chapter summary is stated in Section 6.6.

6.2 Focused Image Compression

In image compression, the tradeoff between the compression ratio and the image quality is hardly controllable automatically [130]. User intervention is often needed to optimize this task. The present section addresses the tradeoff control issue by extending the baseline JPEG algorithm to an adaptive compression method, hereafter adaptive JPEG, which generates compressed images that satisfy both conditions, high compression ratio; and high perceptual quality. Indeed, adaptive JPEG produces automatically-selected regions of interest (ROIs) with a higher reconstruction quality with respect to the rest of the input image, while keeping the overall compression ratio relative high [112]. The ROIs are generated with our purely data-driven visual attention algorithm described in Section 2.4.

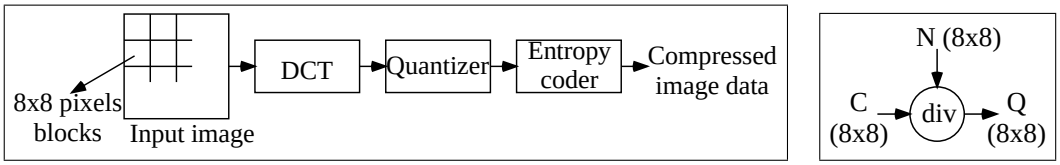
Some previous works have dealt with the problem of the identification of ROIs to spatially adapt the compression according to the relative importance of regions [130, 167]. A more recent work [125, 126] has presented an algorithm based on automatically pre-identified ROIs that have been computed by means of a biologically plausible technique. This technique deals, however, only with grey scale images; chromatic features were not considered.

6.2.1 Baseline JPEG Algorithm

The baseline JPEG algorithm operates on 8×8-pixel blocks $\mathbf{B}$ of an image and is composed of three main steps, as shown in Figure 6.1(a).

1. **DCT**: The Discrete Cosine Transform (DCT) transforms the 8×8 blocks $\mathbf{B}$ from the spatial domain to 8×8 coefficients $\mathbf{C}$ in the frequency domain. This operation aims at removing the data redundancy in the image.

2. **Quantizer**: This operation consists in a rounded division (*div()*) of the 8×8 DCT coefficient matrix $\mathbf{C}$ by a normalization matrix $\mathbf{N}$, which results in a quantized coefficient matrix $\mathbf{Q}$, as shown in Figure 6.1(b). The



(a) Baseline JPEG algorithm (b) Quantizer

Figure 6.1: Baseline JPEG algorithm. (a) represents the different steps of the algorithm, whereas (b) details the operations of the quantizer.

elements of the normalization matrix **N** are chosen so that visually significant DCT coefficients are quantized more accurately than less important coefficients [142].

3. **Entropy coder**: The quantized DCT coefficients are then entropy encoded in order to further compress the data.

For more details about the baseline JPEG algorithm, the reader is referred to [142, 158, 21, 20].

6.2.2 Adaptive JPEG Algorithm

The proposed adaptive JPEG algorithm follows the same operations of the baseline JPEG, albeit with a quantization unit that has been modified to receive an additional input: a scale factor sf, as shown in Figure 6.2. The value of the parameter sf depends on whether a 8×8-pixel block $\mathbf{B}$ lies inside or outside a ROI. Formally, the scale factor sf is set in accordance with Equation 6.1.

$$sf = \begin{cases} sf_0 & if\ card(\mathbf{B} \cap \mathrm{ROI}) \geq \frac{1}{2} \cdot card(\mathbf{B}) \\ sf_1 & otherwise \end{cases} \tag{6.1}$$

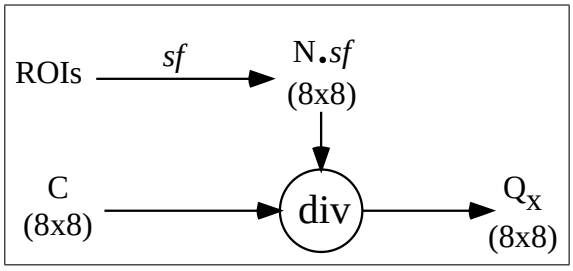


Figure 6.2: Adaptive JPEG algorithm: Quantizer.

Indeed, for those 8x8-pixel blocks with a majority of pixels lying within the ROIs, the quantization is executed using the normalization array previously scaled by sf_0. For the rest of the blocks, **N** is scaled by sf_1 before quantization. To preserve image details within the ROIs, sf_0 is usually chosen to be in

the interval [0.5, 1], while sf_1 is generally selected to be a real number larger than two, which guarantees a high overall compression ratio.

The ROIs that are produced by the visual attention algorithm and used by the quantization unit to adapt the parameter sf, represent overhead information to be embedded in the compressed data bitstream. This data is required for a decoder to execute the corresponding ROI-dependent inverse quantization of the DCT coefficients. This overhead information not only increases the size of the compressed data, but also precludes the compressed image from being reconstructed by a standard JPEG decoder.

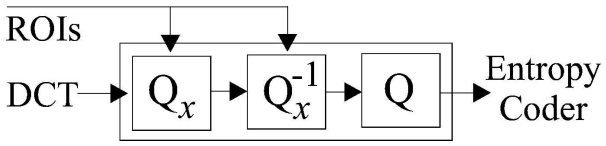


Figure 6.3: Quantizer of the adaptive JPEG algorithm to produce JPEG compatibility.

The problem of the overhead data can be easily overcome in exchange of two additional quantization operations, as shown in Figure 6.3. After the DCT operation, the initial ROI-dependent quantization (Q_x) is followed by a corresponding ROI-dependent inverse quantization (Q_x^{-1}). Indeed the ROI-dependent inverse quantization operation consists in multiplying each 8×8-block of quantized DCT coefficients by sf_0 or by sf_1 depending whether the block lies within or outside a ROI. After this point, the overhead data is no longer required and the current DCT coefficients can be re-normalized using a regular, ROI-independent, JPEG quantization (Q). This procedure produces a spatially adaptive compressed image which is fully compatible with the JPEG standard. This was the scheme used to produce the results presented in Section 6.2.3.

6.2.3 Experimental Results

This section reports experiments involving a set of color images, conducted in order to assess the usefulness of visual attention in the field of color image compression.

For each example, a color image is acquired; using this image as input, the visual attention algorithm computes a set of ROIs considering seven features ($F_{1..7}$) as described in Section 2.4. In these experiments the number of identified ROIs has been, for simplicity, limited to three. Afterwards, the color image is compressed using two methods, a) standard JPEG, and b) Adaptive JPEG. With both methods, and for all experiments, the overall compression ratio produced was 100:1.

The images of the reported experiments are shown in Figure 6.4 and 6.5, they mainly feature two persons facing the camera. Based on the considered features

(chromatic, intensity-based), the two persons stand out from the rest of the scene, and are thus, natural candidates for the ROIs, to be automatically identified by the visual attention algorithm. Figures 6.4(b) and 6.5(b) show, that as expected, the two persons' faces figure among the three most salient regions of the image. The adaptive JPEG algorithm takes into account the relative importance of these

(a) Original image (b) Identified ROIs

(c) Standard JPEG (d) Adaptive JPEG

Figure 6.4: Adaptive versus standard JPEG: Example 1.

image regions. Consequently, the reconstructed images (bottom-right images in Figure 6.4 and 6.5) preserve the visual details of the two faces, which may be relevant for the recognition of the two persons. On the other hand, the persons' faces have lost important perceptual details when using the standard JPEG method (bottom-left images in Figure 6.4 and 6.5). In the latter case, one may have difficulty to identify the two persons. Note that the adaptive JPEG algorithm preserves perceptually important image details at cost of other less significant image information like the background.

The advantage of the adaptive JPEG algorithm is highlighted in Figure 6.6, where the rightmost ROI has been zoomed in.

(a) Original image

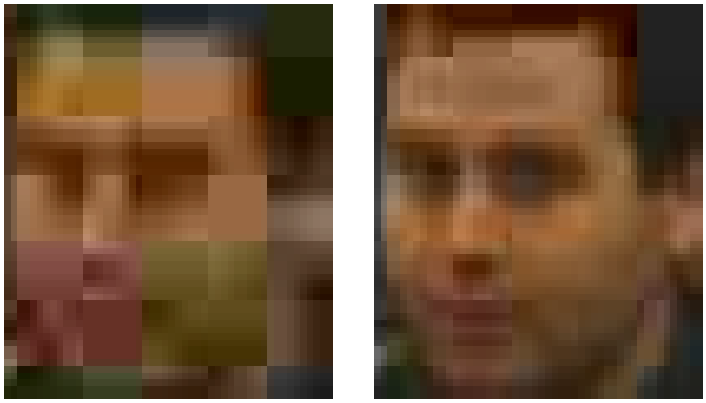
(b) Identified ROIs

(c) Standard JPEG

(d) Adaptive JPEG

Figure 6.5: Adaptive versus standard JPEG: Example 2.

These examples illustrate the capability of the adaptive color image compression algorithm based on visual attention to automatically control the compression/quality tradeoff, despite the unavailability of any a priori knowledge about the analyzed image.



(a) Standard JPEG (b) Adaptive JPEG

Figure 6.6: Adaptive versus standard JPEG: Zoomed ROI from Example 2.

6.3 Attentive Color Image Segmentation

Image segmentation is an essential preprocessing step towards scene understanding in computer vision. The segmentation task aims at grouping together spatially connected pixels which fulfill certain homogeneity criteria. Image segmentation algorithms can be roughly classified into three categories:

(i) **Pixel-based segmentation** is the most local method to address the task of image segmentation. The property of single pixels is used to classify the image points into regions. Histogram thresholding [127] and data clustering [33], among other methods, belong to this category.

(ii) **Edge-based segmentation** relies on discontinuities of image data. The methods belonging to this category are generally composed of three main steps. Firstly, edges are extracted using edge detection techniques [24]. In the second step, non connected edges which belong to the same physical regions border are connected. Finally, regions are derived from the closed edges.

(iii) **Region-based segmentation** is based on two main principles. The first is the feature homogeneity, meaning that pixels of the same region must fulfill certain homogeneity criteria. The other principle is the spatial connectivity of pixels of the same region. Split and merge [29], as well as region growing [1] are classical methods belonging to this category.

Of course there exist also hybrid segmentation methods that combine techniques from different categories to obtain better results [26, 47].

Although built around different concepts, most the segmentation techniques described above have to deal with two major issues related to segmentation, namely the choice of the spatial locations (seeds) where to start the segmentation process and the choice of the homogeneity criteria to segment regions.

Regarding the first issue, numerous segmentation methods use randomly selected seeds to achieve the segmentation task, others use image statistics to determine these seeds. The spots provided by the visual attention algorithm can be seen as natural candidate for the seed points, since they are representative points of salient regions. In addition, focusing the segmentation task on only salient and thus significant regions speeds up not only the segmentation itself but also the subsequent tasks, such as object recognition.

Regarding the second issue numerous segmentation techniques (especially the region-based ones) use the same homogeneity criteria to segment all scene regions. This approach can be seen as a limitation since it neglects the feature-related specificities of single image segments. The new idea is to adapt the homogeneity criteria according to the features that discriminate the region to be segmented from its surroundings, i.e. the salient features.

To address these two issues related to image segmentation, we propose a solution which has been realized in two stages. The first stage of the solution addresses the automatic selection of the image regions which will be considered further for segmentation. Therefore, we have developed a spot-based color image segmentation method which uses the detected spot of attention as seeds to guide a seeded region growing algorithm [110, 111]. In the spot-based method, segmentation is achieved at a fixed scale, using the same homogeneity criteria of regions for all spots.

In a more recent work [115], we extended the spot-based method to achieve segmentation at the appropriate scale and to adapt the homogeneity criteria of regions to the nature of the detected spots. This extension gave rise to MAPS, i.e. Multiscale Attention-based Pre-Segmentation of color images.

6.3.1 Spot-Based Color Image Segmentation

The spot-based color image segmentation that has been reported in [110, 111] relies extensively on the Seeded Region Growing (SRG) algorithm [1], as shown in Figure 6.7. Indeed, our segmentation method starts with computing a set of spots of attention from an input image by means of the saliency-based model of visual attention reported in Section 2.4. Then, these spots are transmitted to a SRG algorithm in order to grow an homogenous image region around each detected spot, according to the following scheme.

Given a multi-channel image $\mathbf{f}$ and a spot of attention P_i, the SRG algorithm, first, initializes a new region R_i with the point P_i. Then, for each neighbor $\mathbf{x}$ of R_i the algorithm checks whether $\mathbf{x}$ satisfies a membership criteria. If it is the case, $\mathbf{x}$ is added to R_i and the mean vector $\overline{\mathbf{f}}_i$ of the region is updated. This

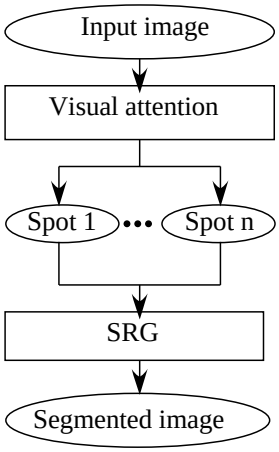


Figure 6.7: Spot-based segmentation algorithm.

procedure is repeated until none of the neighbors of R_i satisfies the membership criteria.

The criteria which a point $\mathbf{x}$ must satisfy in order to be added to a region R_i is given by Equation 6.2.

$$\|\mathbf{f}(\mathbf{x}) - \overline{\mathbf{f}}_i| < \varepsilon \tag{6.2}$$

where $\|.\|$ denotes the euclidian distance.

Formally, the SRG algorithm, applied to a single spot of attention (seed), is described in Algorithm 6.1. The spot-based segmentation consists in sequentially applying the described algorithm for a set of spots of attention.

Algorithm 6.1 Seeded region growing algorithm

```
P_i a spot of attention (seed)
create R_i, initial region from P_i
insert the 8 neighbors of R_i into a list L
while L is not empty do
  remove first point x from L
  if x satisfies the membership criteria in R_i (see Equation 6.2) then
    add x to R_i
    update the mean f̄_i of the region R_i
    insert all neighbors of x into L
  end if
end while
```

Extensive tests of the spot-based segmentation method have been carried out on outdoor RGB images. In these experiments, the spots of attention, which were used as seed points for the segmentation algorithm, have been computed from

seven features; two chromatic features ((R-G), (B-Y)), intensity, and four orientation features. The results[1] demonstrate the significance of the visual attention algorithm for the seed selection procedure that is needed in the segmentation task. As shown in Figure 6.13 (bottom left), the proposed segmentation algorithm has successfully segmented visually salient scene objects (like traffic signs and road boarder line), despite the absence of top-down information about the considered test images. Also, focusing the segmentation on the detected spots of attention has leaded, as expected, to a speed-up of the segmentation task.

6.3.2 MAPS: Multiscale Attention-based Pre-Segmentation of Color Images

In this section, we extend the spot-based algorithm to a Multiscale Attention-based PreSegmentation method (MAPS), which uses featured spots of attention

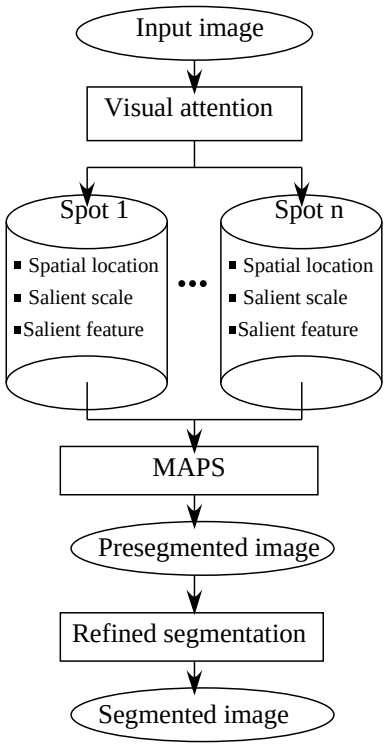


Figure 6.8: MAPS: The saliency-based model of visual attention provides a set featured spots of attention. These data are efficiently used by the segmentation module.

in order to guide the segmentation process [115]. Indeed, each detected spot of

[1] A detailed description of the results is given in [110, 111]

attention is characterized by some features that give insights about the nature of the region to be segmented. The gained information is then efficiently used by the segmentation task. The main steps of MAPS are illustrated in Figure 6.8.

Segmentation-Relevant Scene Data

Before getting into details of the characterization of the detected spots of attention by some segmentation-relevant features, let us recall what the visual attention algorithm provides when applied to a multi-featured image. According to Section 2.4 the data made available by the attention model are listed below.

- J feature maps F_j where j is the feature index.
- $J \cdot K$ multiscale conspicuity maps $\mathcal{M}_{j,k}$ computed at K different scales for the J features (k is the scale index).
- J feature-related conspicuity maps C_j.
- A saliency map $\mathcal{S}$
- A set of spots of attention.

Let us now find out what kind of segmentation-relevant information can be derived from the output of the visual attention algorithm.

As mentioned above and like the spot-based segmentation method described in the previous section, MAPS uses of the spatial locations of salient regions as seeds.

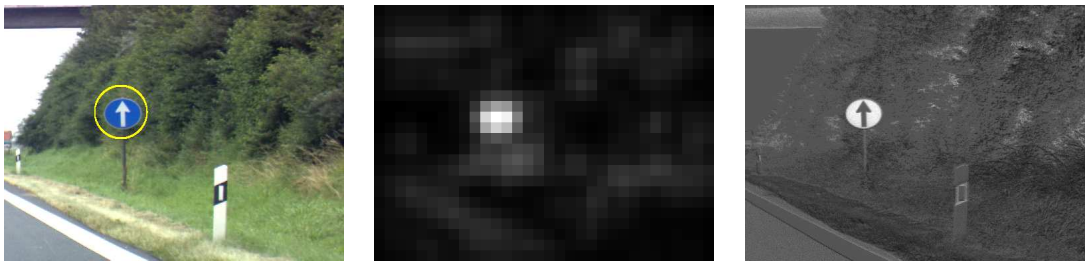
In addition, the salient feature j^* and the salient scale k^* of each detected spot can be determined. Indeed j^* and k^* are the scene feature and the spatial scale that mostly discriminate the detected spot of attention from its surrounding. In order to get these two segmentation-relevant data, we determine, for each spot $\mathbf{x}$, the multiscale conspicuity map $\mathcal{M}_{j^*,k^*}$ (among the $J \cdot K$ maps) which mostly highlights that location. Therefore, (j^*, k^*) are derived according to Equation 6.3.

$$(j^*, k^*) = argmax_{j,k}(M_{j,k}(\mathbf{x})) \tag{6.3}$$

where $M_{j,k} = \mathcal{N}(\mathcal{M}_{j,k})$ is the normalized version of the map $\mathcal{M}_{j,k}$.

To summarize, three kinds of segmentation-relevant information are now available and of which MAPS takes advantage to first achieve the pre-segmentation step and then the refinement step: spatial information (location $\mathbf{x}$ of the detected spots), feature-based information (j^*), and scale-based information (k^*).

Figure 6.9 and Figure 6.12 ((2b) and (3b)) illustrate this presegmentation information which will play an essential role in the segmentation task. Figure 6.9 depicts the salient scale and the map of the salient feature for the first spot of attention on a traffic scene image, whereas Figure 6.12 shows the salient scales and the salient features for the eight spots of attention computed on two traffic scene images.



(a) First spot of attention (b) M_{j^*,k^*} $((j^*,k^*)=(3,3)$ (c) F_{j^*}

Figure 6.9: The segmentation-relevant data provided by MAPS about the first spot of attention on a traffic scene image. $M_{3,3}$ is the multiscale conspicuity map with the highest conspicuity value around this spot. Thus, F_3 (opponent colors B/Y) is the map of the salient feature at this location.

Presegmentation of the Multiscale Map M_{j^*,k^*}

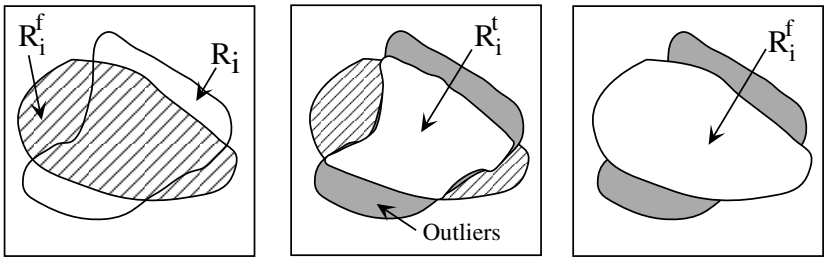
In this section we aim at finding an approximative segmentation (presegmentation) of each detected region, based on their conspicuousness or salience. This presegmentation is best achieved at the salient scale of the considered region. Therefore, we apply, for each detected spot $\mathbf{x}$, the seeded region growing (SRG) algorithm described in Algorithm 6.1, on the corresponding multiscale map $M_{j^*,k^*}(\mathbf{x})$, using the conspicuousness as homogeneity criteria.

Examples of the presegmentation results are illustrate in Figure 6.12 ((2c) and (3c)).

This first step can not be seen as a final segmentation result, since it is achieved at a coarse resolution. Further information collected through the attention model might be used in order to accurately refine the presegmented region at full resolution. This refinement step is best achieved in the salient feature map F_{j^*}.

Refined Segmentation

Given the roughly segmented region R_i and the corresponding feature map F_{j^*}, this procedure consists in refining R_i, which results in the final homogenous (regarding j^*) region R_i^f. Therefore, two main steps are necessary (as illustrated in Figure 6.10): 1) remove the pixels that are merged into R_i and which do not belong to the real region R_i^f (hereafter, outliers). The result of this step is a smaller region R_i^t; 2) insert the pixels that belong to R_i^f but have not been merged into R_i.



(a) Presegmentation (b) Outliers removal (c) Expansion

Figure 6.10: MAPS: Refinement procedure.

Step1: removal of outliers

In order to remove the outliers from R_i, the typical value T of the homogenous region R_i^f is needed. Once T is computed, the outliers removal process is straightforwardly achieved by simple thresholding.

To compute T, we use a statistical method, which consists in the following operations:

1. Compute the histogram $h_{R_i}(m)$ ($m \in [0..255]$) of F_{j^*} within R_i and the corresponding mean value μ_1 according to Equation 6.4.

$$\begin{aligned} h_{R_i}(m) &= \sum_{\mathbf{x} \in R_i | F_{j^*}(\mathbf{x}) = m} (1) \\ \mu_1 &= \frac{1}{card(R_i)} \sum_{\mathbf{x} \in R_i} F_{j^*}(\mathbf{x}) \end{aligned} \tag{6.4}$$

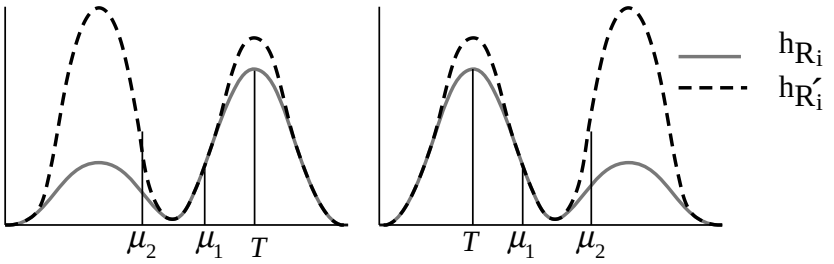
2. Compute the mean value μ_2 of F_{j^*} within a dilated version of the region R_i: $R_i' = dilate(R_i)$ according to Equation 6.5.

$$\mu_2 = \frac{1}{card(R_i')} \sum_{\mathbf{x} \in R_i'} F_{j^*}(\mathbf{x}) \tag{6.5}$$

3. Derive the typical value T in accordance with Equation 6.6.

$$T = \begin{cases} argmax_{(m > \mu_1)}(h_{R_i}(m)) & if\ \mu_1 > \mu_2 \\ argmax_{(m < \mu_1)}(h_{R_i}(m)) & else \end{cases} \tag{6.6}$$

In fact, the order of the two mean values μ_1 and μ_2 determines wether the final region R_i^f is bright with a dark surrounding or the opposite. Once this information is available, we can focus the search of the typical value T of the



(a) Bright region/dark background (b) dark region/Bright background

Figure 6.11: Computation of the typical value T of a presegmented region.

region R_i^f on a small part of the histogram $h_{R_i}(m)$. This idea is graphically illustrated in Figure 6.11.

The removal step ends with a thresholding of R_i using the computed typical value T, which results in a region R_i^t that excludes the outliers, as shown in Figure 6.10(b).

Step2: expand the region

The second and final step of the refinement procedure aims at expanding R_i^t to those pixels that belong to R_i^f and have not been merged into R_i during the presegmentation process. Therefore, the SRG algorithm, described in Algorithm 6.1, is applied to $F_{j^*}(\mathbf{y})$, where $\mathbf{y} = \{\forall \mathbf{x} \in R_i^t\}$.

Segmentation Examples

Figure 6.12 ((2d) and (3d)) illustrates two examples of the refined segmentation on traffic scene color images[2]. For all examples, we have used a visual attention algorithm that considers seven features ($F_{1..7}$ as described in Section 2.4), and that computes six multiscale conspicuity maps for each feature ($K = 6$).

Figure 6.13 depicts a comparison between results achieved by MAPS (Section 6.3.2) and segmentation results of the spot-based segmentation method presented in Section 6.3.1. The comparison reveals two major advantages of MAPS over the spot-based method. On one hand, the mutiscale concept of MAPS permits the segmentation of entire objects rather than local structure of the image (the entire blue traffic signs instead of their white arrows). On the other hand, the presegmentation procedure and the capability of MAPS to adapt the regions homogeneity criteria to the nature of the detected spots allow the segmentation of more significant image regions, such as the segmentation of the sky instead of a part of the forest.

[2]The test images were kindly provided by Klab, Caltech, USA (www.klab.caltech.edu)

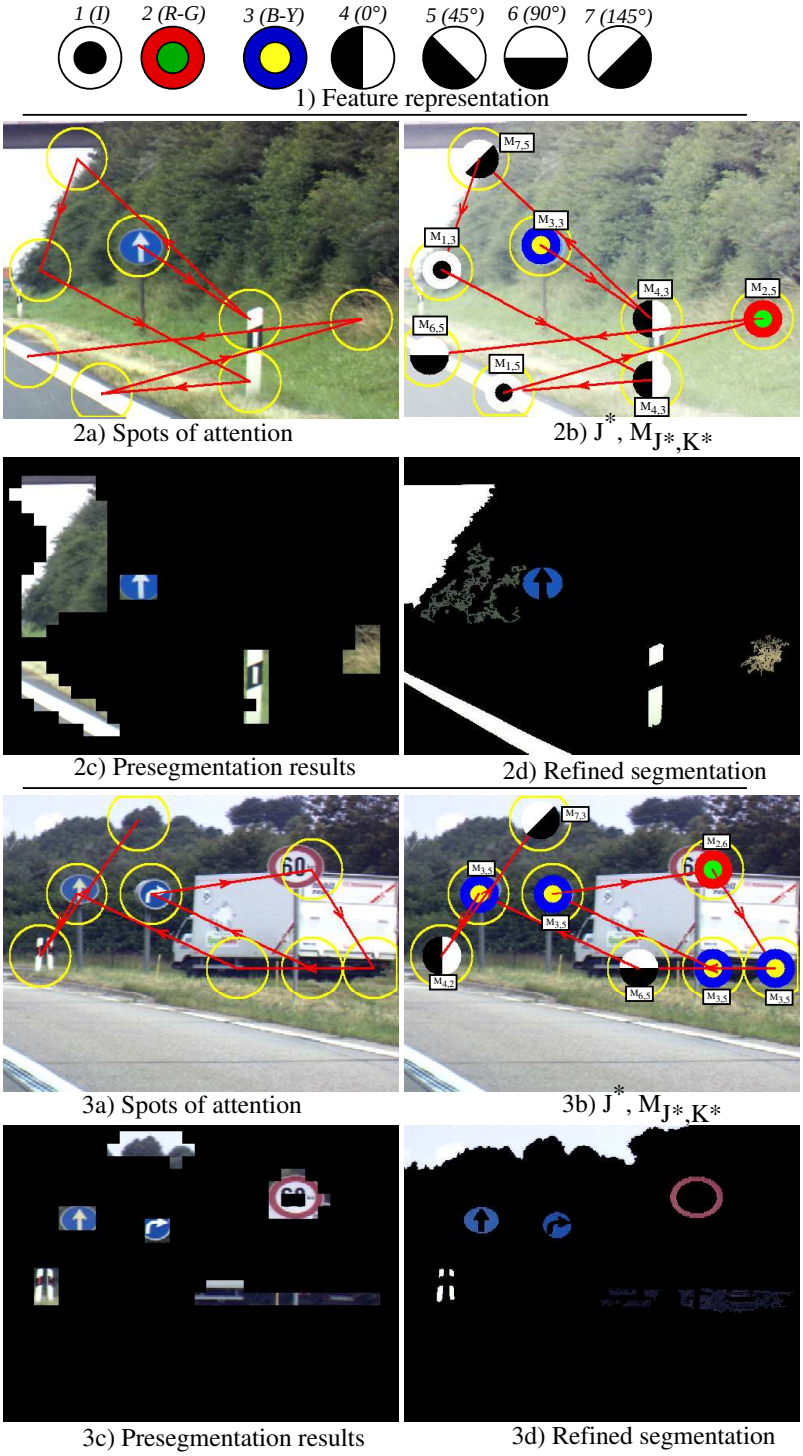
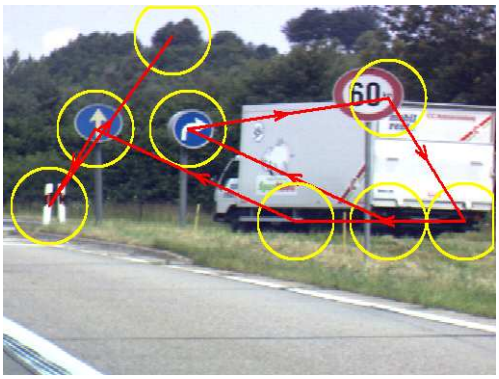


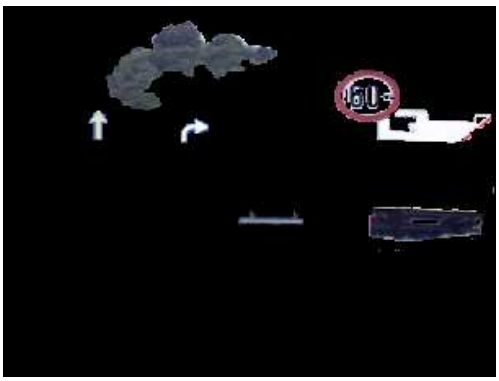
Figure 6.12: Examples of MAPS results. 1): Colored representation of the seven scene features. 2) and 3): Results on two traffic scene images: (a) the spots of attention, (b) corresponding salient features and scales, (c) presegmentation results, and (d) refined segmentation results.

However, MAPS have not been compared to other segmentation methods that also preselect segmentation-relevant scene information statistically, such as clustering.

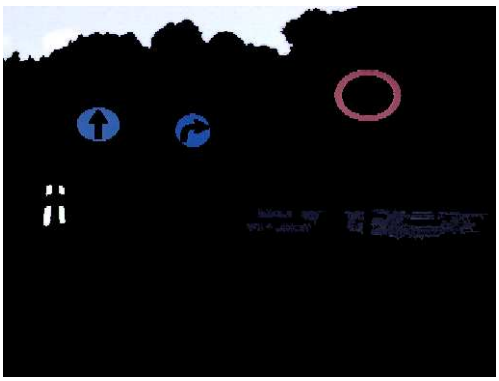
Note that in these experiments only the feature maps $F_{1..3}$ have been used in the refined segmentation procedure. To take full advantage of MAPS, the refined segmentation procedure might be extended, in the future, to consider also salient borders (i.e. salient location originating from orientation conspicuities). A possible realization of such approach is to use contour-based segmentation techniques whenever the orientation represent the salient feature of an image location.



Spots of attention



Spot-based segmentation



MAPS segmentation

Figure 6.13: Spot-based segmentation vs. MAPS segmentation.

6.4 Attention-Based Object Tracking

Given a time sequence of images, object tracking consists in determining the time sequence of spatial positions of specific features. The aim is to pursuit, over time, objects of interest in a dynamic environment, such as video sequences. This task

is of high relevance for numerous applications like video surveillance [32] and visual robot navigation [39].

A robust object tracking can be achieved if two major conditions are fulfilled: 1) object features must be discriminant enough in order to avoid confusion between different objects; 2) object features must be stable over time in order to track them as long as possible. Thus, to properly solve the object tracking problem, two main challenging issues have to be addressed. The first issue is the selection of the objects of interest to be tracked. The second challenge is the characterization of the tracked objects by robust features.

This section reports a novel tracking method, the attention-based tracking [116], which takes advantage of the visual attention paradigm to address the two tracking issues mentioned above. Regarding the selection issue, the model of dynamic visual attention automatically determines, in a dynamic environment, the most salient locations which could be tracked further. The characterization issue is solved by the same model of attention which provides intrinsic and robust features of the detected spots.

Figure 6.14 depicts the principle of the attention-based tracking method.

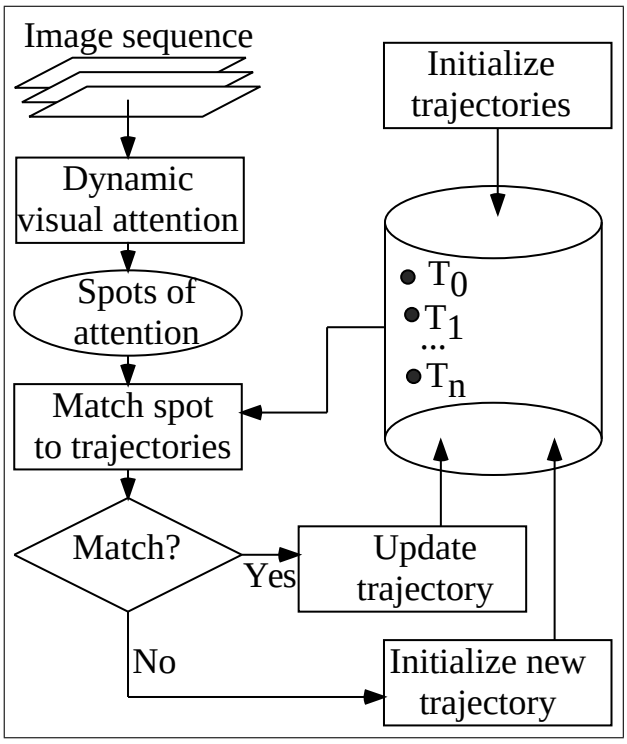


Figure 6.14: Attention-based object tracking.

6.4.1 Object Detection and Characterization

The spots of attention computed from a static conspicuity map C_s and a dynamic one C_d by means of the dynamic model of visual attention (Chapter 3.3) represent the scene objects to be tracked.

Each detected spot $\mathbf{x}$ is characterized by a feature vector $\mathbf{f}$ according to Equation 6.7.

$$\mathbf{f} = \begin{pmatrix} f^m \\ f_1 \\ .. \\ f_J \end{pmatrix} \tag{6.7}$$

where f^m is a motion-related feature which is defined in accordance with Equation 6.8,

$$f^m = \begin{cases} 1 & if\ C_d(\mathbf{x}) > \varepsilon_m \\ 0 & otherwise \end{cases} \tag{6.8}$$

and $f_{1..J}$ are static features which are computed from the corresponding feature-related conspicuity maps C_j as follows:

$$f_j = \frac{\mathcal{N}(C_j(\mathbf{x}))}{C_s(\mathbf{x})} \tag{6.9}$$

Let N be the number of frames of a sequence and M the number of spots detected per frame, the spots of attention can be formally described as $P_{m,n} = (\mathbf{x}_{m,n}, \mathbf{f}_{m,n})$, where $m \in [1..M]$, $n \in [1..N]$, $\mathbf{x}_{m,n}$ is the spatial location of the spot, and $\mathbf{f}_{m,n}$ its characteristic feature vector.

Figure 6.15 gives an example of the detection and characterization of the most salient spots of the test sequence "Munich train station" at frame number 21.

6.4.2 The Tracking Algorithm

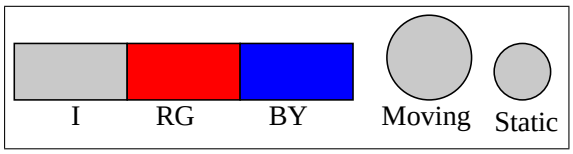
This section presents the algorithm which tracks the characterized spots over time. In conceiving our tracking algorithm, the following assumptions have been made:

- Multiple objects can be tracked, each of which is assigned to a different trajectory T that can be defined as follows:

  $$T = \{P^1, P^2, ..., P^h\} \tag{6.10}$$

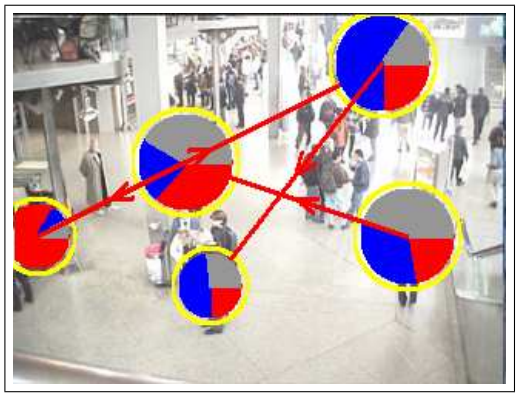
  where $P^h = (\mathbf{x}^h, \mathbf{f}^h)$ is the head element of the trajectory (the least inserted spot).

- Feature constraint: a tracked scene object must have nearly constant characteristic feature vector $\mathbf{f}$ over time. The tolerated feature variation is controlled by a threshold $\varepsilon_{\mathbf{f}}$.



(a) Feature representation

(b) frame 21



(c) Detected and characterized spots

Figure 6.15: Spot characterization by static and dynamic features. Three static features have been considered in this example: intensity (I), red/green color component (RG) and blue/yellow color component (BY) whose color representations are given in (a). The size of the circles encodes the dynamic feature of the objects. (b) represents the frame number 21 of the sequence Munich train station. (c) illustrates the five most salient parts of the scene at that frame and the corresponding characterization.

- Spatial constraint: moving objects are supposed to move within a predictable spatial neighborhood. The largest tolerable displacement $\varepsilon_{\mathbf{x}}$ can be adapted to different situations.

- Temporal constraint: a previously tracked object can disappear from the scene for only a determined period of time in order to be tracked again. This period of time is given by a time-threshold ε_t.

The basic idea behind the proposed algorithm is to build a trajectory for each tracked physical object. Each point of the trajectory memorizes the spatial and the feature-based information of the object at a given time.

Specifically, given the M spots of attention computed from the first frame, the tracking algorithm starts with creating M initial trajectories, each of which contains one of the M initial spots. The corresponding initial spot represents also the head element of the initial trajectory. A new detected spot $P_{m,n}$ is either inserted into an existing trajectory (function $push(P, T)$ in Algorithm 6.2) or gives rise to a new trajectory ($newTraject(T)$), depending on its similarity with the head elements P^h of already existing trajectories. Thus, a trajectory is a list to which new detected spots can be pushed. It is noteworthy that two spots of the same frame n can not be assigned to the same trajectory (see the control parameter $marked[]$ in Algorithm 6.2) and that a spot is assigned to exactly one trajectory.

Formally, the tracking method is given by Algorithm 6.2.

Figure 6.16 illustrates some examples of attention-based object tracking using the image sequence "Munich train station".

6.4.3 Perspectives

We believe that the presented attention-based method of object tracking has large potential in the field of visual robot navigation. In fact the reliability of visual navigation systems extensively relies on the robustness of landmark detection and tracking in unknown environments [44, 39, 162], which makes our tracking method particularly relevant for this application.

The following thoughts can be seen as a basis for the development of a robust robot navigation system.

- Generally, landmarks are conspicuous references of the environment, which makes them easy to detect by our visual attention algorithm.

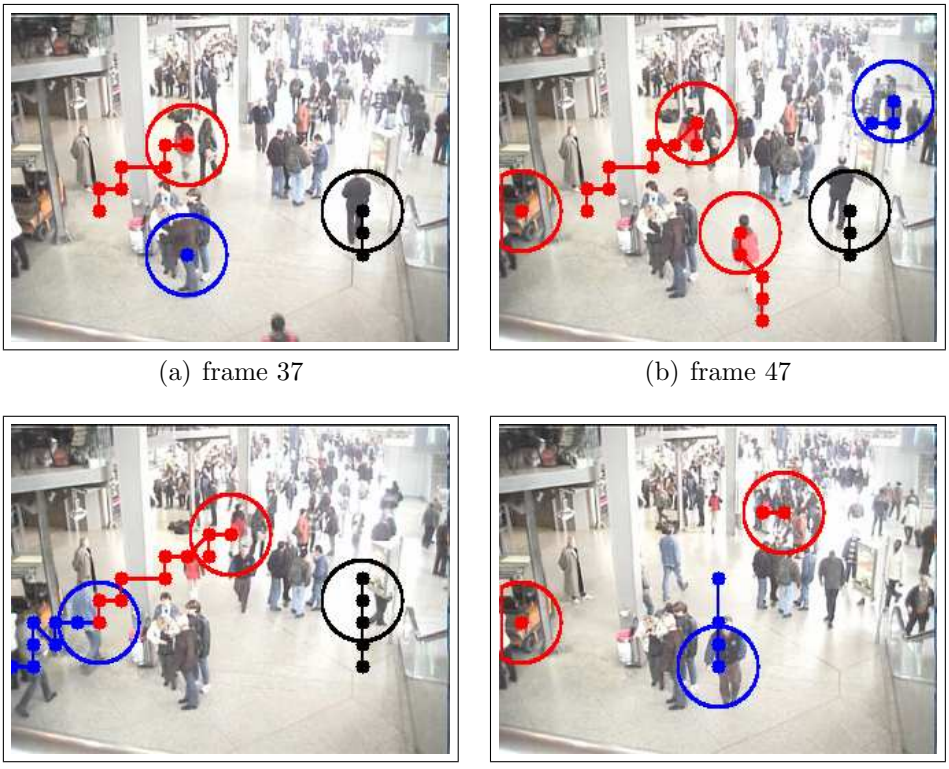
- Landmarks must be tracked over time in order to facilitate the robot positioning and navigation. Our tracking method elegantly solves this task.

- The idea to build trajectories for the tracked objects can be used to further verify the robustness of the detected and tracked landmarks. For instance,

Algorithm 6.2 Attention-based object tracking

Image sequence $I(n)$ $(1..n..N)$
Number of detected spots of attention per frame: M
Boolean $Pushed$
Boolean $marked[\,]$
Trajectory set $\{T\} = \emptyset$

for $n = 1 \,..\, N$ **do**
  Detect & characterize the M spots of attention $P_{m,n} = (\mathbf{x}_{m,n}, \mathbf{f}_{m,n})$
  for $k = 1 \,..\, card(\{T\})$ **do**
    $marked[k] = 0$
  end for
  for $m = 1 \,..\, M$ **do**
    $Pushed = 0$
    for $k = 1 \,..\, card(\{T\})$ **do**
      if $(marked[k] == 0)$ **then**
        if $(\|\mathbf{x}_{m,n} - \mathbf{x}^h\| < \varepsilon_{\mathbf{x}}$ & $\|\mathbf{f}_{m,n} - \mathbf{f}^h\| < \varepsilon_f$ & $|n - n^h| < \varepsilon_t)$ **then**
          $Push(P_{m,n}, T_k)$
          $Pushed = 1$
          $marked[k] = 1$
          break
        end if
      end if
    end for
    if $(Pushed == 0)$ **then**
      $newTraject(T_{card(\{T\})+1})$
      $Push(P_{m,n}, T_{card(\{T\})+1})$
      $\{T\} = \{T\} \cup \{T_{card(\{T\})+1}\}$
    end if
  end for
end for



(a) frame 37 (b) frame 47

(c) frame 66 (d) frame 97

Figure 6.16: Examples of attention-based object tracking.

only persistent scene objects and thus spots that belong to long trajectories can be retained as reliable landmarks of the environment.

6.5 Attention-Based Traffic Sign Recognition System

Vision-based driver assisting systems seem to be a promising solution to increase traffic safety and driving comfort [166, 55]. Automatic traffic sign detection and recognition constitutes an important component of this solution.

In addition to providing a rigorous reliability, an automatic traffic sign detection and recognition system must also operate in real time, since a delay in recognizing a sign could be disastrous to traffic participants. As the previous computer vision applications, this task can use the visual attention paradigm to reach real-time requirements. An attention-based solution for the traffic sign detection problem is particularly promising given the fact that traffic signs are conceived (and also installed) so that the driver perceives them easily. They

almost pop-out from the surrounding landscape.

In a recent diploma thesis work [34], we have developed an attention-based system for traffic sign recognition. The following assumptions have been made during the conception of the system.

- Only a subset of the Swiss traffic signs have been considered, namely triangular red signs, circular red and circular blue signs, as shown in Figure 6.17. Their number amounts to 94 signs.
- The traffic signs are supposed to be not tilted and their frontal side facing the camera.
- The color of the traffic signs is supposed to be clearly perceived (which is not the case for example in rainy or foggy conditions). Indeed, our detection system relies extensively on color-related features.

Figure 6.17: Some examples of the considered traffic signs belonging to three categories: 1) triangular red signs, 2) circular red signs, and 3) circular blue signs.

6.5.1 Overview of the System

The proposed attention-based traffic sign recognition system is composed of two main components (see Figure 6.18): 1) a traffic sign detector which is based on visual attention; 2) a traffic sign recognizer which relies on the data provided by the detector and matches detected regions of interest with traffic sign models.

The visual attention model used for the detection of traffic signs can be seen as an adaptation of the saliency-based model to solve this specific task. Indeed, the features considered in our model are derived from intrinsic characteristics of

the considered traffic signs. Specifically, we used the two color-based features that correspond best to the colors of the considered traffic signs, namely red and blue channels. In addition, triangular and circular shapes, which are intimately linked to the traffic signs are also used as features in our model.

The conspicuity transformation has been also largely adapted to each feature in order to best discriminate the colors or the shape we want to detect. Like in the saliency-based model of visual attention, all conspicuity maps are then combined into the final saliency map, on which the selection of the most salient image locations relies.

The detected spots of attention are used to segment the corresponding regions using the color information [111]. The segmented regions undergoes then a geometric filtering, which filters out those regions which could not correspond to traffic signs using criteria like compactness and elongation.

The regions which pass the geometric filter are finally matched to the traffic sign models stored in a database. Based on the correlation coefficient, the matching method assigns a score to each region/model pair. Depending on this score, the match is either validated and the corresponding traffic sign is recognized or the region is rejected.

6.5.2 Evaluation of the System

This section reports some experiments conducted for evaluating the proposed attention-based traffic sign detection and recognition system. The experiments have been carried out with 138 traffic scene images acquired in various real situations and conditions and containing 182 traffic signs of the categories considered in this work (circular, triangular, red, blue).

The detection modules keeps the five most salient locations. The results of Table 6.1 shows that **85**% of the traffic signs figure amongst them. Regarding recognition, the related module is designed for near zero confusion error at cost of high rejection rate. It recognizes about **55**% of the traffic signs present in the test images. A deep analysis of the recognition module has showed that the segmentation step and the geometric filter, which is highly selective, are responsible for **70**% of the failed recognition.

Figure 6.19 illustrates an example of successful recognition of traffic signs in three different scenes.

To conclude, it can be said that although the recognition module does not reach the expected performance, the detection module seems to be a promising solution for further development of attention-based traffic sign recognition methods.

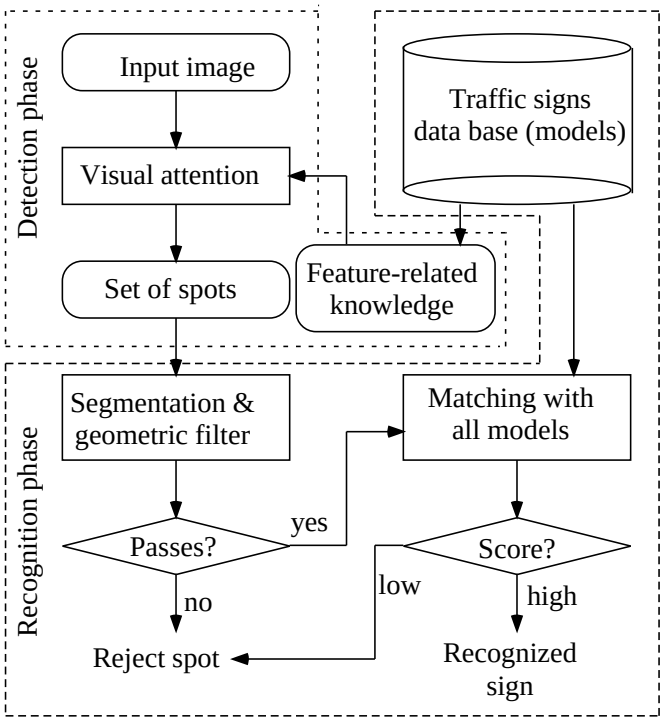


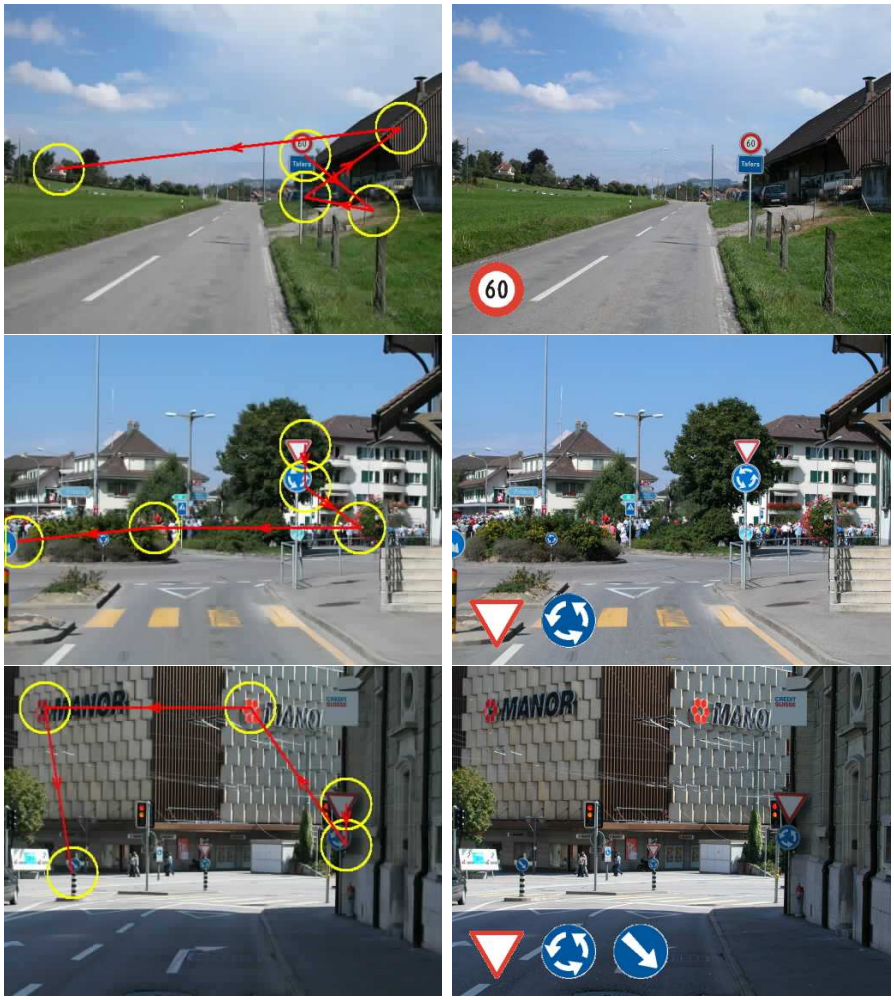
Figure 6.18: Attention-based traffic sign recognition system. First, the proposed system detects regions of interest from a color image by means of a visual attention algorithm whose feature maps are adapted to traffic signs characteristics. Regions around the detected spots are then segmented and a geometric filter is applied on the segmented regions in order to reject those of them which could not correspond to traffic signs. The regions which pass the geometric filter are matched to traffic sign models. Depending on the matching score, the regions are either rejected or assigned a traffic sign.

6.6 Chapter Summary

This chapter has showed the potential of the visual attention paradigm in guiding computer vision applications with the goal, on one hand, to speed them up and on the other, to increase their performance. Indeed, the visual attention algorithm drastically reduces the amount of data to be considered in higher level tasks while preserving the relevant information of a scene.

Four applications related to computer vision have been considered in this context, namely adaptive color image compression, color image segmentation, object tracking, and automatic traffic sign recognition for driver assistance.

Firstly, in the case of adaptive image compression, the visual attention algorithm automatically determines the visually salient image regions which should have a good visual quality after compression and thus for which more bits should



Detected spots Recognized signs

Figure 6.19: Some examples of attention-based traffic sign recognition.

Sign type	Num. of signs	Detected signs	recognized signs
Triangular, red	61	91%	56%
Circular, red	75	84%	51%
Circular, blue	46	79%	58%
All	**182**	**85%**	**55%**

Table 6.1: Evaluation of the traffic sign detector.

be allocated. Hence, the proposed adaptive compression method spatially adapt the compression ratio according to the visual salience of image locations, which represents a potential solution to the problem of automatically controlling the compression ratio/image quality tradeoff.

Second, we have reported MAPS (Multiscale Attention-based PreSegmentation), a color image segmentation method that benefits from the visual attention paradigm regarding two aspects. First, instead of segmenting the entire image, the proposed method considers only salient image parts which leads to a speed up of the segmentation process. Second, the visual attention algorithm provides segmentation-relevant information about each region like salient scale and salient features. The embedding of this region-specific information into the segmentation task increases its performance in comparison to segmentation methods that neglect the specificities of single image regions. Extension of the segmentation method to consider also salient borders is expected to further increase the performance of MAPS.

Third, an attention-based tracking algorithm has been reported. The basic idea is to detect and characterize, by some robust features, the salient locations in the environment. The pre-attentive information is then used to track salient scene objects. It is expected that the proposed object tracking method has a high potential to solve various problems related to visual robot navigation.

Finally, an automatic traffic sign recognition system which takes advantage of the visual attention mechanism has been presented. The system starts with detecting regions that potentially might contain traffic signs. Then, these regions are analyzed in details in order to identify the traffic sign, if any. Although the recognizes only about 55% of the traffic signs, the detection module seems to be a promising solution for further development of attention-based traffic sign recognition methods, since its detection rate reaches 85%.

Chapter 7

Conclusions

Visual attention is an essential component in human vision, since it efficiently solves the trade-off between the amount of transmitted visual information and the processing capacity of the brain. Indeed, visual attention selects the most salient information of a scene and allows the human computation apparatus to allocate its resources for the processing of these attended parts of the scene.

In computer vision, the visual attention paradigm is of high relevance because the computational complexity is a paramount issue, especially if real-time operation is needed. In fact, the selection of only a reduced subset of the image data drastically reduces the complexity of high level computer vision tasks, such as object recognition, which generally has a combinatorial nature. Furthermore, biologically inspired vision mechanisms allow the extraction of robust and relevant features of the scene and contribute to improve the robustness and enhances the overall performance of high level machine vision tasks.

The present thesis proposes new solutions for integrating visual attention in computer vision tasks. The proposed contributions can be classified into three main categories: computational modeling of the visual attention behavior, real-time implementation of a model, and application of the developed concept to solve real computer vision tasks.

For the modeling aspect, the usefulness of novel visual features like depth and motion in visual attention has been assessed. It has been concluded that depth information is of high relevance for visual attention and its complementary nature to 2D features has been demonstrated. In addition, the integration of motion cues strongly influences the behavior of the visual attention model by dramatically increasing the saliency of moving objects in dynamic visual scenes. Furthermore, the plausibility of different versions of the visual attention model has been quantitatively assessed by comparing their attention maps with those obtained for humans, derived from eye movement experiments. More specifically, comparison scores have demonstrated the superiority of the non-linear combina-

tion of scene features into the final attention maps over the linear combination method. The contribution of chromatic features to visual attention has been also quantitatively evaluated and showed to increase by 75% the similarity score existing between human and machine attention maps.

The feasibility of real-time operation of the complex model of visual attention has been also demonstrated in this thesis. Implemented on a general purpose image processing system, consisting of a low power, one board, mixed digital-analog, highly parallel SIMD architecture, the proposed visual attention algorithms operate on 64×64-pixel images at a frequency of 14 Hz, which is enough for processing visual scenes in real-time. Future possible designs should allow a one fold increase in computing performance letting the possibility to integrate additional image features into the visual attention algorithms. Due to its parallel nature, the extension of the architecture to larger images is also possible at the cost of simple hardware replication, provided that the present serial data transfer is upgraded to a parallel form.

The third main contribution of this work is about the application of the visual attention paradigm to solve real computer vision tasks. Four different applications have been considered in this context, namely adaptive color image compression, color image segmentation, object tracking, and automatic traffic sign recognition. In the case of image compression, it has been showed that the visual attention model represents a valuable solution for the automatic control of the compression/quality tradeoff. Concerning segmentation, the most obvious advantage of using visual attention is the overall speed up of the image segmentation procedure. Furthermore, it has been concluded that the performance of this task is clearly enhanced when adding pre-attentive information about image regions. Extension of the proposed method to salient borders would certainly enhance even more the segmentation results. As for object tracking, the application of visual attention opens new perspectives for finding landmarks in visual robot navigation. Automatic traffic signs detection and recognition, with its detection rate of about 85%, provides additional credit to this last statement. To summarize, this work has demonstrated the applicability of the visual attention paradigm to computer vision and hints to the high potential of attention-based computer vision solutions.

The work presented in this thesis opens many perspectives in several challenging research fields. Possible extensions of our work include:

- Empirical validation on human subjects of the 3D visual attention model. Indeed, the recent availability of stereoscopic displays makes possible the recording of human visual attention behavior for 3D scenes, which should allow the assessing of the plausibility of our 3D model of attention. Pre-

liminary results are very encouraging.

- Empirical validation on human subjects of the dynamic model of visual attention. Such an extension constitutes a research field where novel results are expected, since, to our knowledge, no publications exist on this subject.
- Integration of a top-down component into the saliency-based model of visual attention. Basically, the saliency-based model of visual attention is purely bottom-up and thus lacks a systematic consideration of a priori-knowledge about scenes. The extension of the visual attention model with a top-down component can take model on previous works, e.g. [96, 164, 89].
- Applying visual attention to visual navigation of autonomous mobile systems, like mobile robots. Indeed, the preliminary results we have recently obtained in this field are promising.

Based on the most important issues covered in this thesis, it can be concluded that this thesis work has given rise to a multi-cue biologically plausible model of visual attention that can operate in real-time and whose applicability to computer vision was demonstrated. This work has also opened numerous perspectives in conceiving and implementing bio-inspired computer vision solutions.

Appendix A

Itti's Implementation

A.1 Multiscale Conspicuity Maps

Given the resolution pyramid $\mathcal{P}_j(.)$, six multiscale conspicuity maps $\mathcal{M}_{j,k}$ are computed for a feature j, in accordance with Equation A.1.

$$\begin{aligned}
\mathcal{M}_{j,1} &= |\mathcal{P}_j(2) - \mathcal{P}_j(5)| \\
\mathcal{M}_{j,2} &= |\mathcal{P}_j(2) - \mathcal{P}_j(6)| \\
\mathcal{M}_{j,3} &= |\mathcal{P}_j(3) - \mathcal{P}_j(6)| \\
\mathcal{M}_{j,4} &= |\mathcal{P}_j(3) - \mathcal{P}_j(7)| \\
\mathcal{M}_{j,5} &= |\mathcal{P}_j(4) - \mathcal{P}_j(7)| \\
\mathcal{M}_{j,6} &= |\mathcal{P}_j(4) - \mathcal{P}_j(8)|
\end{aligned} \tag{A.1}$$

A.2 Gabor Pyramids

For a given orientation θ, the Gabor pyramid (also known as orientation pyramid) $\mathcal{O}_\theta()$ is built according the following scheme.

1. Build Laplacian pyramid $\mathcal{L}()$ as follows:

$$\mathcal{L}(n) = \mathcal{P}(n) - \mathcal{P}(n+1) \tag{A.2}$$

2. For each level n of the Laplacian pyramid, modulate the laplacian image with a sine wave and apply a lowpass filter (LPF) on the result. Let $\mathcal{L}'(n)$ be the resulting image:

$$\mathcal{L}'(n) = LPF[e^{(i\vec{k}_\theta \cdot \vec{r})} \cdot \mathcal{L}(n)] \tag{A.3}$$

where $\vec{r} = x\vec{i} + y\vec{j}$($x$ and y are the spatial coordinates of the Laplacian image), and $\vec{k}_\theta = (\pi/2)[\cos\theta\vec{i} + \sin\theta\vec{j}]$. Note that $\mathcal{L}'(n)$ is complex valued.

3. Finally, the orientation pyramid or the Gabor pyramid is computed as the absolute value (modulus) of $\mathcal{L}'$. Indeed, for each level n, $\mathcal{O}_\theta()$ is given by Equation A.4.

$$\mathcal{O}_\theta(n) = |\mathcal{L}'(n)| \tag{A.4}$$

Appendix B

Gradient-Based Optical Flow

B.1 Gradient Constraint Equation

Let $I(\mathbf{x}, t) = I(x, y, t)$ be the image function at time t, and $\delta\mathbf{x} = (\delta x, \delta y)$ the displacement of an image point (x, y) within a time period of δt. At time $t + \delta t$ the image function is $I(\mathbf{x}+\delta\mathbf{x}, t+\delta t)$. A first order Taylor expansion of this term leads to Equation B.1.

$$I(\mathbf{x} + \delta\mathbf{x}, t + \delta t) = I(\mathbf{x}, t) + \frac{\partial I}{\partial x}\delta x \ + \frac{\partial I}{\partial y}\delta y \ + \ \frac{\partial I}{\partial t}\delta t \tag{B.1}$$

Moving $I(\mathbf{x}, t)$ to the left side of Equation B.1, and divide both sides by δt gives Equation B.2.

$$\frac{I(\mathbf{x} + \delta\mathbf{x}, t + \delta t) - I(\mathbf{x}, t)}{\delta t} = \frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} \tag{B.2}$$

According to the "brightness conservation" assumption, we have:
$I(\mathbf{x}+\delta\mathbf{x}, t+\delta t) = I(\mathbf{x}, t)$ (see Equation 3.7). If we combine this information and Equation B.2, we obtain Equation B.3.

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0 \tag{B.3}$$

Now, $(\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t})$ is nothing else than the optical flow $\mathbf{v} = (u, v)$. Thus, Equation B.3 can be rewritten into Equation B.4.

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{B.4}$$

Using vectorial representation, Equation B.4 can be transformed into Equation B.5.

$$\left[\begin{array}{cc} \frac{\partial I}{\partial x} & \frac{\partial I}{\partial y} \end{array}\right] \cdot \left[\begin{array}{c} u \\ v \end{array}\right] + \frac{\partial I}{\partial t} = 0 \tag{B.5}$$

Using the appropriate annotations (e.g. I_t instead of $\frac{\partial I}{\partial t}$), the gradient constraint equation is trivially deduced (Equation B.6).

$$\nabla I(\mathbf{x}, t) \; . \; \mathbf{v} + I_t(\mathbf{x}, t) = 0 \tag{B.6}$$

Bibliography

[1] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 16, No. 6*, 1994.

[2] S. Ahmed. *VISIT: An Efficient Computational Model of Human Visual Attention.* PhD thesis, University of Illinois at Urbana-Champaign, 1991.

[3] J.Y.. Aloimonos. Purposive and qualitative active vision. *International Conference on Pattern Recognition ICPR'90, pp. 346-360*, 1990.

[4] J.Y. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *International Journal on Computer Vision, Vol. 1, pp. 333-356*, 1987.

[5] A.I.E. Alpaydin. *Incremental Learning and Selective Attention in a Biologically-inspired Model of Character Recognition.* PhD thesis, EPFL, Department of Computer Science, Switzerland, 1990.

[6] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision, Vol. 2, pp. 283-310*, 1989.

[7] L. Arendes. *Aufmerksamkeitseffekte in Zellen des Superioren Colliculus bei Makaken.* PhD thesis, University of Goettingen, Shaker Press, Aachen, 1993.

[8] X. Arreguit, F.A. Van Schaik, F.V. Bauduin, M. Bidiville, and E. Raeber. A CMOS motion detector system for pointing devices. *IEEE Journal of Solid State Circuits, Vol. 31, No. 12, pp. 1916-1921*, 1996.

[9] N. Ayache and F. Lustman. Fast and reliable passive trinocular stereovision. *International Conference on Computer Vision ICCV'87, pp. 422-427*, 1987.

[10] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 23, No. 12, pp. 1415-1429*, 2001.

[11] J.L. Barron, S.S. Beauchemin, and D.J. Fleet. On optical flow. *Bioimaging, Vol. 2, No. 1, pp. 57-61*, 1994.

[12] J.L. Barron, A.D. Jepson, and J.K. Tsotsos. The feasibility of motion and structure from noisy time-varying image velocity information. *International Journal of Computer Vision, Vol. 5, No. 3, pp. 239-269*, 1990.

[13] R Battiti, E. Amaldi, and Ch. Koch. Computing optical-flow across multiple scales - an adaptive coarse-to-fine strategy. *International Journal of Computer Vision, Vol. 6, pp. 133-145*, 1991.

[14] T.M. Bernard, B.Y. Zavidovique, and F.J. Devos. A programmable artificial retina. *IEEE Journal of Solid State Circuits, Vol. 28, No. 7, pp. 789-797*, 1993.

[15] P. Besl. Geometric signal processing. *In RC Jain and AK Jain (Eds), Springer-Verlag*, 1990.

[16] M. Bollmann. *Entwicklung einer Aufmerksamkeitssteuerung fuer ein aktives Sehsystem.* PhD thesis, Department of Computer Science, University of Hamburg, 1999.

[17] M. Bollmann, C Justkowski, and B. Mertsching. Utilizing color information for the gaze control of an active vision system. *4th. Workshop der Farbbildverarbeitung, pp. 73-79*, 1998.

[18] M. Boucart, A.M. Henaff, and C. Belin. Vision: aspects perceptifs et cognitifs. *Edition Solal*, 1998.

[19] P. Bouthemy and E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision, Vol. 10, No. 2, pp. 159-182*, 1993.

[20] J. Bracamonte. *Algorithms, VLSI Architectures, and ASIC design for Image Compression Systems.* PhD thesis, University of Neuchâtel, Switzerland, 1998.

[21] J. Bracamonte, M. Ansorge, and F. Pellandini. Adaptive block-size transform coding for image compression. *ICASSP'97, Vol. 4, pp. 2721-2724*, 1997.

[22] V. Brajovic and T. Kanade. Computational sensor for visual tracking with attention. *IEEE Journal of Solid State Circuits, Vol. 33, No. 8, pp. 1199-1207*, 1998.

[23] K. Brunnstrom, J.O. Eklundh, and T. Uhlin. Active fixation for scene exploration. *International Journal of Computer Vision, Vol. 17, pp. 137-162*, 1994.

[24] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 8, pp. 679-698*, 1986.

[25] K.R. Cave and J.M. Wolfe. Modeling the role of parallel processing in visual search. *Cognitive Psychology, Vol. 22, pp. 255-271*, 1990.

[26] A. Chakraborty and J.S. Duncan. Game-theoretic integration for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 21, No. 1, pp. 12-30*, 1999.

[27] Y. Chan, M. Yu and A. Constantinides. Variable size block matching motion compensation with application to video coding. *Proceeding of IEE-I, Vol. 137, No. 4, pp. 205-212*, 1990.

[28] D. Chapman. *Vision, Instruction and Action.* PhD thesis, AI Laboratory, Massachusetts Institute of technology, 1990.

[29] S.Y. Chen, W.C. Lin, and C.T. Chen. Split and merge image segmentation based on localized feature analysis and statistical tests. *CVGIP, Vol. 53, pp. 457-475*, 1991.

[30] G.T. Chou. A model of figure-ground segregation from kinetic occlusion. *International Conference on Computer Vision, pp. 1050-1057*, 1995.

[31] J.J. Clark and N.J. Ferrier. Control of visual attention in mobile robots. *IEEE Conference on Robotics and Automation, pp. 826-831*, 1989.

[32] I. Cohen and G. Medioni. Detecting and tracking objects in video surveillance. *IEEE Conference on Computer Vision and Pattern Recognition, pp. 2319-2325*, 1999.

[33] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. *Computer Vision and Pattern Recognition. pp. 750-755*, 1997.

[34] O. Corbat. Reconnaissance automatique des signaux routiers pour véhicules intelligents. *Rapport de travail de diplôme, Insititut de microtechnique, université de Neuchâtel*, 2003.

[35] J.L. Crowley and H.I. Christensen. *Vision as process.* Springer Verlag, 1993.

[36] CSEM:. Centre suisse d'electronique et microtechnique. *http://www.csem.ch/.*

[37] S.M. Culhane and J.K. Tsotsos. An attentional prototype for early vision. *2nd European Conference on Computer Vision, Lecture Notes in Computer Science, Springer Verlag, Vol. 588, pp. 551-560*, 1992.

[38] S.M. Culhane and J.K. Tsotsos. A prototype for data-driven visual attention. *International Conference on Pattern Recognition (ICPR), Vol. 1, pp. 36-40*, 1992.

[39] A.J. Davison. *Mobile Robot Navigation Using Active Vision.* PhD thesis, University of Oxford, UK, 1999.

[40] E. De Micheli, V. Torre, and S. Uras. The accuracy of the computation of optical flow and of the recovery of motion parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) Vol. 15, No. 5, pp. 434-447*, 1993.

[41] R. Deriche. Fast algorithms for low-level vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 12, No. 1, pp. 78-87*, 1990.

[42] R. Desimone, M. Wessinger, L. Thomas, and W. Schneider. Attentional control of visual perception: cortical and subcortical mechanisms. *In: Cold Spring Harbor on Quantitative Biology, Vol. LV: The Brain, Cold Spring Harbor Laboratory Press, pp. 963-971*, 1990.

[43] S. Dickinson, H.I. Christensen, J.K. Tsotsos, and G. Olofsson. Active object recognition integrating attention and viewpoint control. *CVIU, Vol. 67, No. 3, pp. 239-260*, 1997.

[44] J.A. Driscoll, R.A. Peters, and K.R. Cave. A visual attention network for a humanoid robot. *International Conference on Intelligent Robotic Systems*, 1998.

[45] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature, Vol. 388, No. 6637, pp. 68-71*, 1997.

[46] C.W. Eriksen and Y.Y. Yeh. Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performances, Vol. 11, pp. 583-597*, 1985.

[47] J. Fan, D.K.Y. Yau, A.K. Elmagarmid, and W.G. Aref. Automatic image segmentation by integrating color edge extraction and seeded region growing. *IEEE Transactions on Image Processing, Vol. 10, No. 10, pp. 1454-1466*, 2001.

[48] O.D. Faugeras. *Three-dimensional computer vision: a geometric viewpoint.* MIT Press, 1993.

[49] O.D. Faugeras and L. Robert. What can two images tell us about a third one?. *International Journal of Computer Vision, Vol. 18, pp. 5-20*, 1996.

[50] J.M. Findlay. Saccade target selection during visual search. *Vision Research, Vol. 37, pp. 617-631*, 1997.

[51] J.M. Findlay and I.D. Gilchrist. *Eye Guidance and Visual Search.* In G. Underwood (Ed.), Eye Guidance in Reading and Scene Perception, Oxford, Elsevier Science Ltd, pp. 295-312, 1997.

[52] J.M. Findlay and R. Walker. A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences, Vol. 22, pp. 661-721*, 1999.

[53] B. Fischer and H. Weber. Express saccades and visual attention. *Behavioral and brain science, Vol. 16, No. 3*, 1993.

[54] D.J. Fleet and A.D. Jepson. Computation of component image velocity from local phase information. *International Journal on Computer Vision, Vol. 5, pp. 77-104*, 1990.

[55] D.M. Gavrila. Traffic sign recognition revisited. *21st DAGM Symposium fuer Mustererkennung, Informatik aktuell, Springer Verlag, pp. 86-93*, 1999.

[56] M.S. Gazzaniga. *The New Cognitive Neurosciences.* MIT Press, Cambridge, Massachusetts, 1999.

[57] S. Gil and R. Milanese. Combining multiple motion estimates for vehicle tracking. *Computer Vision - ECCV 96, pp. 307-320*, 1996.

[58] M.S. Gizzi, E. Katz, R.A. Schumer, and J.A. Movshon. Selectivity for orientation and direction of motion of single neurons in cat striate and extratriate visual cortex. *Journal of Neurophysiology, Vol. 63, pp. 1529-1543*, 1990.

[59] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C.H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 222-228*, 1994.

[60] R. Groner and M.T. Groner. Attention and eye movement control: An overview. *European Archives of Psychiatry and Neurological Sciences, Vol. 239, pp. 9-16*, 1989.

[61] D. P. Hanes, W. F. Patterson, and J. D. Schall. The role of frontal eye field in countermanding saccades: Visual, movement and fixation activity. *Journal of Neurophysiology, Vol. 79, pp. 817-834*, 1998.

[62] D.J. Heeger. Optical flow using spatio-temporal filters. *International Journal on Computer Vision, Vol. 1, pp. 279-302*, 1988.

[63] D. Heinke and G.W. Humphreys. Computational models of visual selective attention: A review. *In Houghton, G., editor, Connectionist Models in Psychology*, in press.

[64] SH. Hendry and C. Reid. The koniocellular pathway in primate vision. *Annual Review Neuroscience, Vol. 23, pp. 127-153*, 2000.

[65] J.E. Hoffman and B. Subramaniam. Saccadic eye movements and visual selective attention. *Perception and Psychophysics, 57, pp. 787-795*, 1995.

[66] B.K.P. Horn and B. Schmuck. Determining optical flow. *Artificial Intelligence, Vol. 17, pp. 185-203*, 1981.

[67] G. Indiveri. Modeling selective attention using a neuromorphic VLSI device. *Neural Computation, 2000. Vol. 12, No. 12, pp. 2857-2880*, 2000.

[68] L. Itti. Bottom-up visual attention home page. *http://ilab.usc.edu/bu/*, 1998.

[69] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention.* PhD thesis, California Institute of Technology, USA, 2000.

[70] L. Itti. Real-time high-performance attention focusing in outdoors color video streams. *SPIE Human Vision and Electronic Imaging IV (HVEI'02), pp. 235-243*, 2002.

[71] L. Itti and Ch. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. *SPIE Human Vision and Electronic Imaging IV (HVEI'99), Vol. 3644, pp. 373-382*, 1999.

[72] L. Itti and Ch. Koch. Learning to detect salient objects in natural scenes using visual attention. *Image Understanding Workshop, pp. 1201-1206*, 1999.

[73] L. Itti and Ch. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203*, 2001.

[74] L. Itti, Ch. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259*, 1998.

[75] R.J.K. Jacob. Eye tracking in advanced interface design. *in W. Barfield & T. Furness, eds, Advanced Interface Design and Virtual Environments, Oxford University Press, pp. 258-288*, 1995.

[76] B. Julesz and J. Bergen. Textons, the fundamental elements in preattentive vision and perception of textures. *Bell System Technical Journal, Vol. 62, No. 6, pp. 1619-1645*, 1983.

[77] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 16, No. 9, pp. 920-932*, 1994.

[78] E.M. Klier, H. Wang, and J.D. Crawford. The superior colliculus encodes gaze commands in retinal coordinates. *Nature Neurosceience, Vol. 4, No. 6, pp. 627-632*, 2001.

[79] Ch. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, Vol. 4, pp. 219-227*, 1985.

[80] Ch. Koch, H.T. Wang, R. Battiti, B. Mathur, and C. Ziombowski. An adaptative multi-scale approach for estimating optical flow: computational theory and physiological implementation. *IEEE Workshop on Visual MotionPrinceton, pp. 111-123*, 1991.

[81] T. Komuro and M. Ishikawa. 64 x 64 pixels general purpose digital vision chip. *11th IFIP International Conference on Very Large Scale Integration (VLSI-SOC'01), Montpellier, pp. 327-332*, 2001.

[82] A. A. Kustov and D.L. Robinson. Shared neural control of attentional shifts and eye movements. *Nature, Vol. 384, pp. 74-77*, 1997.

[83] D. LaBerge. Attention, awareness, and the triangular circuit. *Consciousness and Cognition, Vol. 6, pp. 149-181*, 1997.

[84] D. LaBerge, M. Carter, and V. Brown. A network simulation of thalamic circuit operations in selective attention. *Neural Computation, Vol. 4, pp. 318-331*, 1992.

[85] A.G. Leventhal. The neural basis of visual function. *Vision and visual dysfunction, Boca Raton, FL:CRC Press, Vol. 4*, 1991.

[86] M.D. Levine. Vision in man and machine. *McGraw-Hill*, 1985.

[87] T. Lineberg and L. Bretzner. Real-time scale selection in hybrid multi-scale representations. *4th International Conference on Scale-Space theories in Computer Vision, Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2695, pp. 148-163*, 2003.

[88] N.K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology Vol. 5, pp. 552-563*, 1995.

[89] M. Lopez, M.A. Fernandez, A. Fernandez-Caballero, and A. Delgado. Neurally inspired mechanisms of the dynamic visual attention map generation task. *7th International Work Conference on Artificial and Natural Neural Networks, IWANN2003, Lecture Notes in Computer Science, Springer-Verlag, Vol. 2686, pp. 694-701*, 2003.

[90] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *DAPRA IU Workshop, pp. 121-130*, 1981.

[91] C. Maioli, I. Benaglio, S. Siri, K. Sosta, and S. Cappa. The integration of parallel and serial processing mechanisms in visual search: Evidence from eye movement recording. *European Journal of Neuroscience, Vol. 13, pp. 364-372*, 2001.

[92] A. Maki and J.O. Eklundh. A computational model of depth-based attention. *ICPR*, 1996.

[93] A. Maki, P. Nordlund, and J.O. Eklundh. Attentional scene segmentation: Integrating depth and motion from phase. *Computer Vision and Image Understanding, Vol. 78, pp. 351-373*, 2000.

[94] S.K. Mannan, K.H. Ruddock, and D.S. Wooding. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, Vol. 10, No. 3, pp. 165-188*, 1996.

[95] D. Marr. *Vision*. Freeman Publishers, 1982.

[96] R. Milanese. *Detecting Salient Regions in an Image: from Biological Evidence to Computer implementation*. PhD thesis, Dept. of Computer Science, University of Geneva, Switzerland, 1993.

[97] R. Milanese, T. Pun, and H. Wechsler. A non linear integration process for the selection of visual information. *Intelligent Perceptual Systems: New Directions in Computational Perception, Lecture Notes in Artificial Intelligence, Springer Verlag, Vol. 745, pp. 322-336*, 1993.

[98] A. Moini. *Vision chips*. Kluwer, 2000.

[99] M. Mozer and M. Sitton. Computational modeling of spatial attention. *In H. Pashler (Ed.), Attention, London: UCL Press, pp. 341-393*, 1996.

[100] M.C. Mozer. *The perception of multiple objects: a connectionist approach.* MIT Press, 1991.

[101] H.H. Nagel. On the estimation of optical flow: Relations between different approaches and some new results. *Artificial Intelligence, Vol. 33, pp. 299-324*, 1987.

[102] U. Neisser. Cognitive psychlogy. *Appleton, New York*, 1967.

[103] E. Niebur, L. Itti, and Ch. Koch. Modeling visual selective attention: The 'where' pathway. *Models of Neural Networks, Springer Verlag, pp. 247-276*, 2002.

[104] E. Niebur and Ch. Koch. Control of selective visual attention: Modeling the *where* pathway. *Advances in Neural Information Processing Systems, Vol. 8, pp. 802-808*, 1996.

[105] E. Niebur and Ch. Koch. Computational architectures for attention. *The Attentive Brain.R Parasuraman, R., ed., . MIT Press, Cambridge, Massachusetts, pp. 163-186*, 1998.

[106] Center of Psychology University of Athabasca. Tutorial 24: Brain visual pathways. *http://psych.athabascau.ca/html/Psych402/Biotutorials/*, 1997.

[107] R. Okada, Y. Shirai, and J. Miura. Object tracking based on optical flow and depth. *IEEE/SICE/RSJ International Conference on MFI, pp. 565-571*, 1996.

[108] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 15, No. 4, pp. 353-363*, 1993.

[109] B. Olshausen, C.H. Anderson, and D.C. Van Essen. A neural model of visual attention and invariant pattern recognition. *California Institute of Technology, Computation and Neural System Program, CNS Memo 18*, 1992.

[110] N. Ouerhani, N. Archip, H. Hugli, and P. J. Erard. Visual attention guided seed selection for color image segmentation. *International Conference on Computer Analysis of Images and Patterns (CAIP'01), Springer Verlag, LNCS 2124, pp. 630-637*, 2001.

[111] N. Ouerhani, N. Archip, H. Hugli, and P. J. Erard. A color image segmentation method based on seeded region growing and visual attention. *International Journal of Image Processing and Communication, Vol. 8, No. 1, pp. 3-11*, 2002.

[112] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini. Adaptive color image compression based on visual attention. *International Conference on Image Analysis and Processing (ICIAP'01), IEEE Computer Society Press, pp. 416-421*, 2001.

[113] N. Ouerhani and H. Hugli. Attention visuelle: une application qui se base sur la perception visuelle multimodale. *Rapport IMT, université de Neuchâtel*, 1999.

[114] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. *International Conference on Pattern Recognition (ICPR'00), IEEE Computer Society Press, Vol. 1, pp. 375-378*, 2000.

[115] N. Ouerhani and H. Hugli. Maps: Multiscale attention-based presegmentation of color images. *4th International Conference on Scale-Space theories in Computer Vision, Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2695, pp. 537-549*, 2003.

[116] N. Ouerhani and H. Hugli. A model of dynamic visual attention for object tracking in natural image sequences. *International Conference on Artificial and Natural Neural Network (IWANN), Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2686, pp. 702-709*, 2003.

[117] N. Ouerhani and H. Hugli. Real-time visual attention on a massively parallel SIMD architecture. *International Journal of Real Time Imaging, Vol. 9, No. 3, pp. 189-196*, 2003.

[118] N. Ouerhani, H. Hugli, P.Y. Burgi, and P.F. Ruedi. A real time implementation of visual attention on a SIMD architecture. *DAGM 2002, Springer Verlag, Lecture Notes in Computer Science (LNCS), Vol. 2449, pp. 282-289*, 2002.

[119] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Mueri. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis (ELCVIA), Vol. 3 (1), pp. 13-24*, 2004.

[120] F. Paillet, D. Mercier, and T.M Bernard. Second generation programmable artificial retina. *IEEE ASIC/SOC Conference*, 1999.

[121] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research, Vol. 42, No. 1, pp. 107-123*, 2002.

[122] Perception and Eye Movements Laboratory. University department of neurology, Inselspital, Bern, Switzerland. *http://www.neuro-bern.ch/oculo/*.

[123] M.I. Posner and J. Fan. Attention as an organ system. *To appear in J. Pomerantz editor Neurobiology of Perception and Communication: From Synapse to Society the IVth De Lange Conference. Cambridge UK:Cambridge University Press*, in press.

[124] M.I. Posner and S.E. Petersen. The attention system in human brain. *Annual Review of Neuroscience, Vol. 13, pp. 25-42*, 1990.

[125] C. Privitera and L. Stark. Focused JPEG encoding based upon automatic preidentified regions of interests. *SPIE, Vol. 3, pp. 552-558*, 1999.

[126] C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *Pattern Analysis and Machine Intelligence (PAMI), Vol. 22, No. 9, pp. 970-981*, 2000.

[127] J. Puzicha, T. Hofmann, and J. Buhmann. Histogram clustering for unsupervised image segmentation. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99), pp. 602-608*, 1999.

[128] Point Grey Research. Color triclops. *http://www.ptgrey.com/*, 1999.

[129] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience Vol. 2, pp. 1019-1025*, 1999.

[130] R. Rosenholtz and A.B. Watson. Perceptual adaptive JPEG coding. *IEEE International Conference on Image Processing, Vol. 1, pp. 901-904*, 1996.

[131] P.-F. Ruedi, P.R. Marchal, and X. Arreguit. A mixed digital-analog SIMD chip tailored for image perception. *International Conference on Image Processing 96, Vol. 2, pp. 1011-1014*, 1996.

[132] A. Salah, E. Alpaydin, and L. Akrun. A selective attention based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 3, pp. 420-425*, 2002.

[133] D.D. Salvucci. A model of eye movements and visual attention. *Third International Conference on Cognitive Modeling, pp. 252-259*, 2000.

[134] R. Sekuler and R. Blake. Perception. *McGraw-Hill (2nd edition)*, 1990.

[135] M. Sheperd, J.M. Findlay, and R.J. Hockey. The relationship between eye movements and spatial attention. *Journal of experimental psychology, Vol. 38, pp. 475-491*, 1996.

[136] E. Simoncelli. Coarse-to-fine estimation of visual motion. *Eighth Workshop on Image and Multidimensional Signal Processing*, 1993.

[137] E. Simoncelli. *Distributed Analysis and Representation of Visual Motion.* PhD thesis, Massachusetts Institute of Technology, 1993.

[138] A. Singh. An estimation-theoretic framework for image flow computation. *International Conference on Computer Vision, pp. 168-177*, 1990.

[139] D.L. Standley. An object position and orientation IC with embedded imager. *IEEE Journal of Solid State Circuits, Vol. 26, No. 12, pp. 1853-1859*, 1991.

[140] L. Stark and S. Ellis. Scanpaths revisited: Cognitive models direct active looking. *In Eye Movements: Cognition and Visual Perception, edited by D. F. Fisher, R. A. Monty and J. W. Senders. Hillside, NJ: L. Erlbaum Associates*, 1981.

[141] C. Sun. Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. *International Journal of Computer Vision, Vol. 47, pp. 99-117*, 2002.

[142] ITU-T Recommendation T.81. Digital compression and coding of continuous-tone still images. 1992.

[143] K.G. Thompson and J. D. Schall. The detection of visual signals by macaque frontal eye field during masking. *Nature Neuroscience, Vol. 2, pp. 283-288*, 1999.

[144] S.J. Thorpe. Traitement d'images chez l'homme. *Techniques et sciences informatiques, Vol. 7, No. 6, pp. 517-525*, 1988.

[145] S.J. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature, Vol. 81, No. 6582, pp. 520-522*, 1996.

[146] E. Todt and C. Torras. Detection of natural landmarks through multi-scale opponent features. *ICPR 2000, Vol. 3, pp. 988-1001*, 2000.

[147] A.M. Treisman. Spreading suppression or feature integration? a reply to duncan and humphreys. *Journal of Experimental Psychology: Human Perception and Performance, Vol. 18, No. 2, pp. 589-593*, 1992.

[148] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology, pp. 97-136*, 1980.

[149] A.M. Treisman and S. Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, Vol. 95, pp. 15-48*, 1988.

[150] J.K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Science, Vol. 13, pp. 423-469*, 1990.

[151] J.K. Tsotsos. Toward computational model of visual attention. *In T. V. Papathomas, C. Chubb, A. Gorea & E. Kowler, Early vision and beyond, MIT Press, pp. 207-226*, 1995.

[152] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence, Vol. 78, pp. 507-545*, 1995.

[153] J.K. Tsotsos, M. Pomplun, Y. Liu, J.C. Martinez-Trujillo, and E. Simine. Attending to motion: Localizing and classifying motion patterns in image sequences. *Second International Workshop on Biologically Motivated Computer Vision (BMCV'02), pp. 439-452*, 2002.

[154] S. Uras, F. Girosi, A. Verri, and V. Torre. A computational approach to motion perception. *Biological Cybernetics, Vol. 60, pp. 79-97*, 1988.

[155] D.C. Van Essen and C.H. Anderson. Information processing strategies and pathways in the primate retina and visual cortex. *Introduction to Neural and Electronic Networs, Academic Press, pp. 43-72*, 1990.

[156] E.A. Vittoz and X. Arreguit. Linear networks based on transistors. *Electronic Letters, Vol. 29, pp. 297-299*, 1993.

[157] R. von Wartburg, N. Ouerhani, R. Mueri, H. Hugli, and C.W. Hess. Methods for the empirical validation of a computational model of visual attention. *Joint Meeting: Swiss Society for Neuroscience (SSN) and Swiss Society of Psychiatry and Psychotherapy (SSPP)*, 2002.

[158] G.K. Wallace. The jpeg still picture compression standard. *Communications of the ACM, pp. 30-45*, 1991.

[159] T.N. Walther, L. Itti, M. Riesenhuber, T. Poggio, and Ch. Koch. Attentional selection for object recognition - a gentle way. *2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), pp. 472-479*, 2002.

[160] T. Watanabe *et al.* Attention-regulated activity in human primary visual cortex. *Journal of Neurophysiology, Vol. 79, pp. 2218-2221*, 1998.

[161] T. Watanabe *et al.* Task-dependent influences of attention on the activation of human primary visual cortex. *National Academy for Science, Vol. 95, pp. 11489-11492*, 1998.

[162] C.J. Westelius. *Focus of Attention and Gaze Control for Robot Vision.* PhD thesis, Linkoping University, Sweden, 1995.

[163] N. Winters and J. Santos-Victor. Visual attention-based robot navigation using information sampling. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'01), pp. 1670-1675*, 2001.

[164] J.M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, Vol. 1, pp. 202-238*, 1994.

[165] A.L. Yarbus. Eye movements and vision. *New York: Plenum Press*, 1967.

[166] M.M. Zadeh, T. Kasvand, and C.Y. Suen. Localization and recognition of traffic signs for automated vehicle control systems. *Conference on Intelligent Transportation Systems, part of SPIE's Intelligent Systems & Automated Manufacturing, pp. 272-282*, 1997.

[167] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, and Y. Matsuschita. A JPEG codec adaptive to the relative importance of regions in an image. *Transaction of Information Processing Society of Japan Vol. 38, No. 8, pp. 1531-1542*, 1997.