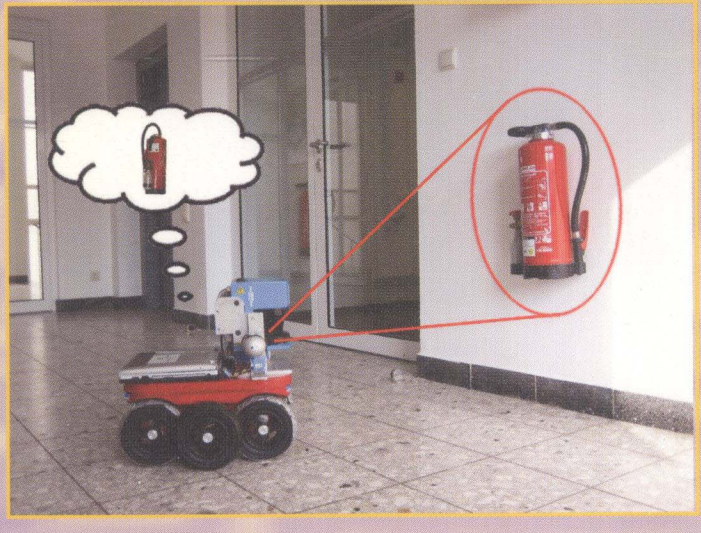
Simone Frintrop

VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search

LNAI 3899



Springer

Simone Frintrop

VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Author

Simone Frintrop
Kungliga Tekniska Högskolan (KTH)
Computer Science and Communication (CSC)
Computational Vision and Active Perception Laboratory (CVAP)
10044 Stockholm, Sweden
E-mail: frintrop@csc.kth.se, simone.frintrop@web.de

This work was carried out at
Fraunhofer Institute for Autonomous Intelligent Systems (AIS)
St. Augustin, Germany
and accepted as PhD thesis at the University of Bonn, Germany

Library of Congress Control Number: 2006921341

CR Subject Classification (1998): I.2.10, I.2.6, I.4, I.5, F.2.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-540-32759-2 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-32759-2 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11682110 06/3142 5 4 3 2 1 0

Foreword

In humans, more than 30% of the brain is devoted to visual processing to allow us to interpret and behave intelligently as part of our daily lives. Vision is by far one of the most versatile and important sensory modalities for our interaction with the surrounding world. Consequently, it is not surprising that there is a considerable interest in endowing artificial systems with similar capabilities. Computational vision for embodied cognitive agents offers important competencies in terms of navigating in everyday environments, recognition of objects for interaction and interpretation of human actions as part of cooperative interaction.

One problem in terms of use of vision is computational complexity. It is well known that tasks such as search and recognition in principle might have NP complexity. At the same time, for use of vision in natural environments there is a need to operate in real-time, and thus to bound computational complexity to ensure timely response. The study of visual attention is very much the design of control mechanisms to limit complexity. Using a rather coarse classification one might divide visual processing into data- and model/goal-driven processing. In data-driven processing, the areas of an image to be processed are selected based on their saliency and offered to other modules in a system for higher-level tasks as, for example, recognition and description. So this is very much the "What is out there?" type of processing. In model-driven processing, the processing is driven by a desire to answer questions such as "Is there a cup in the image?". The selection of which regions to process and how to fuse different image descriptors is then performed according to criteria of optimality in the sense of discrimination.

Visual attention has been widely studied for at least a century, and over the last 25 years rich models of visual attention in primates have been developed. This is not to say that a complete model is available; in fact, a number of competing models have been reported in the literature. However, there are well-formulated models from biology which can be adopted for computational systems.

and, in loving memory, to my father. Both have always believed in me and permanently supported me in every way. Without their help, love, and faith I never would have been able to even start this work.

Abstract

Visual attention is a mechanism in human perception which selects relevant regions from a scene and provides these regions for higher-level processing as object recognition. This enables humans to act effectively in their environment despite the complexity of perceivable sensor data. Computational vision systems face the same problem as humans: there is a large amount of information to be processed and to achieve this efficiently, maybe even in real-time for robotic applications, the order in which a scene is investigated must be determined in an intelligent way. A promising approach is to use computational attention systems that simulate human visual attention.

This monograph introduces the biologically motivated computational attention system VOCUS (Visual Object detection with a CompUtational attention System) that detects regions of interest in images. It operates in two modes, in an exploration mode in which no task is provided, and in a search mode with a specified target. In exploration mode, regions of interest are defined by strong contrasts (e.g., color or intensity contrasts) and by the uniqueness of a feature. For example, a black sheep is salient in a flock of white sheep. In search mode, the system uses previously learned information about a target object to bias the saliency computations with respect to the target. In various experiments, it is shown that the target is on average found with less than three fixations, that usually less than five training images suffice to learn the target information, and that the system is mostly robust with regard to viewpoint changes and illumination variances.

Furthermore, we demonstrate how VOCUS profits from additional sensor data: we apply the system to depth and reflectance data from a 3D laser scanner and show the advantages that the laser modes provide. By fusing the data of both modes, we demonstrate how the system is able to consider distinct object properties and how the flexibility of the system increases by considering different data. Finally, the regions of interest provided by VOCUS serve as input to a classifier that recognizes the object in the detected region. We show how and in which cases the classification is sped up and how the detection quality is improved by the attentional front-end. This approach is

especially useful if many object classes have to be considered, a frequently occurring situation in robotics.

VOCUS provides a powerful approach to improve existing vision systems by concentrating computational resources to regions that are more likely to contain relevant information. The more the complexity and power of vision systems increase in the future, the more they will profit from an attentional front-end like VOCUS.

Acknowledgments

First, I would like to express my profound gratitude to my advisor, Prof. Joachim Hertzberg, who supported my work with many valuable hints and suggestions and who always took the time to answer my questions and to comment on my writings. I was also deeply impressed by his skills and I am indebted to him for enabling me to study the here presented subject in depth; without this, I never would have been able to finish this thesis so rapidly. Special thanks go to Erich Rome, who supported this work with many useful suggestions and who was available each time I asked for his help. I am also grateful to Prof. Armin B. Cremers, who kindly took on the task to co-advise this thesis. Furthermore, Prof. Wolfgang Förstner's valuable suggestions, which helped me to seriously improve my work, are greatly appreciated. I was impressed by his bright scientific mind and by his strong enthusiasm for science.

I am also deeply grateful to Prof. John Tsotsos for his kind advice. He contributed to my work with very helpful ideas and always kept me going; I particularly enjoyed our inspiring e-mail discussions. Furthermore, I want to thank Gerriet Backer for the fruitful discussions on many aspects of this thesis. His helpful comments regarding the psychological background on attention and suggestions concerning computational realizations contributed considerably to my work.

I also would like to thank all my colleagues for supporting me in various ways, especially Andreas Nüchter, Kai Pervölz, Matthias Hennig, Sara Mitri, Uwe Weddige, and Hartmut Surmann, for the fruitful collaboration and the pleasant working atmosphere. Several people kindly provided me with their image data or experimental results. Special thanks go to Jens Pannekamp, Bernd Schönwälder, Fred Hamker, Vidhya Navalpakkam, and Laurent Itti.

Finally, I want to sincerely thank Henrik for his enduring patience when I started to discuss my topic after work or at weekends, for showing interest in my work, and for constantly encouraging me and cheering me up during the tough times. I am also grateful to my friends with whom I had an enjoyable life beyond work. Last but not least, my very special thanks go to my mother

The present volume is an excellent example of how such computational models can be adopted for artificial systems and how we can study these models empirically using robots. Simone Frintrop has chosen to base her research on the popular model by Koch and Ullman, which is based on the psychological work by Treisman termed the "feature-integration-theory". The model uses saliency maps in combination with a winner-take-all selection mechanism. Once a region has been selected for processing, it is inhibited to enable other regions to compete for the available resources. The Koch-Ullman model has primarily been studied for data-driven/bottom-up processing. The framework presented in the present volume — the VOCUS (Visual Object Detection with a Computational Attention System) — presents a modification of the Koch-Ullman model to enable both data-driven and model-driven integration of features. Through adaption of a hybrid model it is possible to integrate the control strategies for search and recognition into a single attentional mechanism.

VOCUS includes a strategy for direct learning of object models for later recognition. It is well suited for design of artificial systems to be used in application, for example, in cognitive systems or in robotics. The volume contains not only a basic design of the hybrid attention model, but the new method has also been tested on detection and recognition of objects in everyday scenarios such as indoor office navigation and recognition of objects on a cluttered tabletop. VOCUS has in addition been evaluated for detection of objects using laser range data, which represents an extreme version of a dense disparity field. Using such diverse sets of feature representations, highly efficient strategies for both search and recognition have been reported.

Simone Frintrop has thus achieved significant progress on several fronts. First of all, the new model represents a major step forward on integration of data and model-driven mechanisms for studies of visual attention. In addition, the model has been empirically evaluated using a diverse set of visual scenes to clearly characterize the new model. It is highly encouraging to see this synthesis of earlier results from primate attention work into a joint model and to see the application of the attention model in the context of robotic applications for navigation and scene modeling.

Henrik I. Christensen
Stockholm, December 2005.

List of Acronyms

CODE	COntour DEtector theory for perceptual grouping
CTVA	CODE Theory of Visual Attention
DAM	Distributed Associative Memory
FEF	Frontal Eye Fields
FIT	Feature Integration Theory
fMRI	functional Magnetic Resonance Imaging
FOA	Focus Of Attention
IOR	Inhibition Of Return
IPL	Inferior Parietal Lobule
IT	Infero Temporat cortex
LGN	Lateral Geniculate Nucleus
LIP	Lateral IntraParietal area
MFG	Middle Frontal Gyrus
MSR	Most Salient Region
MT	Middle Temporal area (V5)
NE	NorepinEphrine system
NVT	Neuromorphic Vision Toolkitt
PO	Parieto Occipale area
PP	Posterior Parietal cortex
ROI	Region Of Interest
RT	Reaction Time
SC	Superior Colliculus
SPL	Superior Parietal Lobule
SAIM	Selective Attention for Identification Model
SEF	Supplementary Eye Field
SERR	SEarch via Recursive Rejection
SLAM	SeLective Attention Model
TVA	Theory of Visual Attention
V1	primary visual cortex, striate cortex
V2 - V5	regions of extrastriate cortex
VOCUS	Visual Object detection with a CompUtational attention System
WTA	Winner Take All network

Contents

1

Introduction

1.1 Motivation

Imagine the following scenario: you are visiting the street carnival in Cologne, Germany for the first time. Fascinated by the colorful and imaginative costumes of the people around you, your gaze wanders from one exciting spot to the next: here a clown with a fancy dress, there a small boy masqueraded as Harry Potter. But not only visual cues capture your attention: over there a band starts to play the new hit of the year and the smell of fresh cookies from the right also revives your interest. Suddenly you remember that you did not come here alone: where has your friend gone? You start to look around, finding her is not easy in the crowd. You remember that she wears a yellow hat, a clue that could make the search easier and you start to watch out for yellow hats. After your gaze has been distracted by some other yellow spots, you detect the hat, recognize your friend who is just dancing with a group of witches, and you start to push through the crowd to join them.

This scenario gives an insight into the complexity of human perception. A wealth of information is perceived at each moment, much more than can be processed efficiently by the brain. Nevertheless, detection and recognition of objects usually succeed with little conscious effort. In contrast, in computer vision and robotics the detection and recognition of objects is one of the hardest problems [Forsyth and Ponce, 2003]. There are several sophisticated systems for specialized tasks such as the detection of faces [Viola and Jones, 2004] or pedestrians [Papageorgiou et al., 1998] – although even these approaches usually fail if the target is not viewed frontally – but developing a general system able to match the human ability to recognize thousands of objects from different viewpoints, under changing illumination conditions and with partial occlusions seems to lie remotely in the future. Suggesting therefore to improve the performance of technical systems is to seek for inspiration from biological systems and to simulate their mechanisms – the brain is the proof that solving the task is possible.

One of the mechanisms that make humans so effective in acting in everyday life is the ability to extract the relevant information at an early processing stage, a mechanism called *selective attention*. The extracted information is then directed to higher brain areas where complex processes such as object recognition take place. Restricting these processes to a limited subset of the sensory data enables efficient processing.

One of the main questions when determining the *relevant* information is the problem of what is relevant. There is no general answer since the relevance of information depends on the situation. With no special goal except exploring the environment, certain cues with strong contrasts attract our attention, for example the clown in the fancy dress. The saliency also depends on the surrounding: the clown is much more salient in a crowd of black witches than among other clowns. In addition to these *bottom-up cues*, the attention is also influenced by *top-down cues*, that means cues from higher brain areas like knowledge, motivations and emotions. For example, if you are hungry the smell of fresh cookies might capture your attention and cause you to ignore the clown. Even more demanding is a goal: when you start to search for the yellow hat of your friend you concentrate on yellow things on the heads of the people around you. Other cues, even if salient, lose importance. Both bottom-up and top-down cues compete for attention and direct your gaze to the most interesting region. The choice of this region is not only based on visual cues but, as suggested in the carnival example, sounds, smells, tactile sensations, and tastes also compete for attention.

In computer vision and robotics, object detection and recognition is a field of high interest. Applications in computer vision range from video surveillance, traffic monitoring, driver assistance systems, and industrial inspection to human computer interaction, image retrieval in digital libraries and medical image analysis. In robotics, the detection of obstacles, the manipulation of objects, the creation of semantic maps, and the detection of landmarks for navigation profit considerably from object recognition.

The further the development of such systems proceeds and the more general their tasks will be, the more urgent is the need for a pre-selecting system that sorts out the bulk of irrelevant information and helps to concentrate on the currently relevant data. A system that meets these requirements is the visual attention system VOCUS (Visual Object detection with a CompUtational attention System) that will be presented in this work.

1.2 Scope

In this monograph, a computational attention system, VOCUS, is presented, which detects regions of potential interest in images. First, fast and rough mechanisms compute saliencies according to different features like intensity, color, and orientation in parallel. If target information is available, the features are weighted according to the properties of the target. Second, the resulting

information is fused and the most salient region is determined, yielding the focus of attention. Finally, the focus region is provided for complex processes like object recognition, which are usually costly and time consuming. By restricting the complex tasks to small portions of the input data, the system is able to achieve considerable performance gains.

The introductory example presented above already contains the four main aspects of the monograph which are examined in the four main chapters: first, VOCUS detects regions of interest from bottom-up cues such as strong contrasts and uniqueness (e.g., the fancy clown); second, top-down influences such as goal-dependent properties influence the processing and enable goal-directed search (e.g., the yellow hat); third, information from different sensor modes attracts the attention and is fused to yield a single focus of attention (as the music and the smell of cookies compete for attention with the visual cues) and finally, after directing the focus of attention to a region of interest, object recognition takes place (e.g., recognition of the hat).

Now some words to categorize the present work. There are two objectives usually aspired by computational attention systems. The first is to better understand human perception and provide a tool that is able to test whether the psychological models are plausible. The second objective is to build a technical system which represents a useful front-end for higher-level tasks as object recognition and thus assists to yield a faster and more robust recognition system. This monograph concentrates on the second objective, that means the aim of the work is to build a system that improves the recognition performance in computer vision and robotics.

1.3 Contributions

This monograph presents a new approach for robust object detection and goal-directed search in images. The work is based on a well-known and widely accepted bottom-up attention system [Itti et al., 1998]. This architecture is extended and improved in several aspects, the major one being extending the system to deal with top-down influences and perform goal-directed search. A detailed discussion on the delimitation to existing work follows in the respective chapters, here we present a short summary of the main contributions:

- Introduction of the computational attention system VOCUS which extends and improves one of the standard approaches of computational attention systems [Itti et al., 1998] by several aspects, ranging from implementation details to conceptual revisions. These improvements enable a considerable gain in performance and robustness (chapter 4, also published in [Mitri et al., 2005, Frintrop et al., 2005c, Frintrop et al., 2005b]).
- Presentation of a new top-down extension of VOCUS to enable goal-directed search. Learning of target-specific properties as well as searching for the target in a test scene are performed by the same attention system. Detailed experiments and evaluations of the method illustrate the

behavior of the system and demonstrate its robustness in various settings. This is the main contribution of the monograph (chapter 5, also published in [Frintrop et al., 2005a, Mitri et al., 2005, Frintrop et al., 2005b]).
- Extension of the attention model to enable operation on different sensor modes. Application of the system to range and reflection data from a 3D laser scanner and investigation of the advantages of the respective sensor modes (chapter 6, also published in [Frintrop et al., 2005c, Frintrop et al., 2003a, Frintrop et al., 2003b]).
- Combination of the attention system with a classifier that enables object recognition. Evaluation of the time and quality performance that is achieved by combining the systems (chapter 7, also published in [Frintrop et al., 2004b, Frintrop et al., 2004a, Mitri et al., 2005]).

Several aspects of these contributions have been done in cooperation with some of my colleagues: the data acquisition with the laser scanner (chapter 6 and 7) has been performed by Andreas Nüchter and Hartmut Surmann. The object recognition with the classifier (chapter 7) has been done in cooperation with Andreas Nüchter, Sara Mitri and Kai Pervölz. Some of the experiments concerning goal-directed search (chapter 5) have been performed by Uwe Weddige. Furthermore, many valuable hints and suggestions were given by Joachim Hertzberg, Erich Rome, and Gerriet Backer.

1.4 Outline

The remainder of this monograph is structured into six chapters. The first two are concerned with the psychological and neuro-scientific background of visual attention (chapter 2) and with the state of the art of computational attention systems (chapter 3), whereas the following four chapters each deal with one of the main contributions of this work:

Chapter 4 introduces the computational attention system describing the details that enable the computation of a region of interest. Particular emphasis is placed on the improvements with respect to other systems and on the discussion of how bottom-up systems of attention may be evaluated.

Chapter 5 elaborates on top-down influences as a new approach to bias the processing of visual input according to the properties of a target object. It is shown how these properties are learned from one or a small selection of training images, and how the learned information is used to find the target in a test scene. A wide variety of experiments on artificial as well as on real-world scenes show the effectiveness of the system.

Chapter 6 examines the extension of VOCUS to several sensor modes. The application of the attention system to range and reflection data from a 3D laser scanner illustrates how the information may be processed separately and finally fused into a combined representation from which a single focus of attention is computed. The advantages of each sensor mode are discussed and the differences between saliencies in laser and camera data are highlighted.

Chapter 7 combines the attention system with a fast and powerful classifier to enable recognition on the region of interest. It is shown how the time and quality performance improves when combining the two systems. Finally, chapter 8 concludes the work by summarizing the main concepts, discussing the strengths and limitations, and giving an outlook on future work.