Lecture Notes in Artificial Intelligence 3899

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

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

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
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.

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.

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

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.

Contents

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