TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Kognitive Systeme
Fakultät für Elektro- und Informationstechnik

Enabling Scalable and Efficient Visual Attention, Object-Based Attention and Object Recognition for Humanoid Robots - a Biologically-Inspired Approach.

Dipl.-Inf. Univ. Andreas Holzbach

Vollständiger Abdruck der von der Fakultät für Elektro- und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Ingenieur-Wissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. Werner Hemmert

Prüfer der Dissertation:
1. Univ.-Prof. Gordon Cheng, Ph.D.
2. Prof. Dr.-Ing. Aleš Ude, Univ. Ljubljana/Slowenien

Die Dissertation wurde am 16.04.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektro- und Informationstechnik am 15.04.2016 angenommen.

ABSTRACT

As easy as it seems for us to visually perceive and recognize our environment, we are still just at the beginning of understanding how the human vision system might work in its whole. Complex neuroscientific models have been built, drawing their foundation both from neurophysiology, as well as empirical-obtained data from psychological studies. Most of the existing models aim at representing and evaluating different theorems, disregarding their real-world applicability - being in terms of efficiency, robustness and generality. Pure technical-oriented vision systems however are confined to quite specific scenarios - like face recognition or tracking of objects - and seem to fail to function in a more general context while humans or animals still outperform any technical system by far.

This thesis presents an approach towards the efficient integration of neuroscientific knowledge into a technical environment for improving vision models in time-crucial real-world scenarios in the context of humanoid robotics. Instead of copying and computationally replicating what we know about the information processing in the brain, this thesis tries to grasp the pure functional aspect of it, in order to enhance technical systems.

The main contribution of this thesis is the analysis, the development and the evaluation of an efficient biologically-inspired vision system which supplies a humanoid robot with the ability to visually perceive and understand its environment in an efficient and scalable manner. The proposed model integrates three essential parts of human vision: Visual attention, object-based attention and object recognition. Visual and object-based attention play a major role in how and what we perceive in our field of vision by selecting and reducing the available information. Both of which are essential for a fast and reactive vision system.

Six major contributions of this thesis helped to build this model: a) the design of a new visual attention system, which outperforms state-of-the-art system in terms of ac-

curacy, speed and complexity realized by b) our Sampled Template Collation method for efficiently evaluating different image regions and which is able to adapt to computational needs; c) a new object-based attention system, which enhances object recognition; d) our object recognition model - an enhancement of a computational model called HMAX, which is an abstraction of the neural information processing in the visual cortex. The model was modified in terms of speed and performance in order to put it in a more technical context. We quantitatively and qualitatively show that by the integration of neuroscientific knowledge about neural information processing in the brain like lateral-inhibition and avoidance of entropic redundancy results in a higher classification accuracy and faster processing speed; e) the development of a temporal reasoning framework which enables the system to classify over time and account for uncertainties in non-static real-world scenarios; f) the system architecture for the efficient integration of visual attention, object-based attention and object recognition, which enabled the humanoid robot iCub to detect and segment even fast moving objects like a thrown ball.

The whole system was realized using a cluster architecture with multi-core CPUs and GPUs to spread the computational payload and match the strong concurrent and highly parallel character of information processing in the brain. Further enhancement involved the application of methods for optimization from signal detection theory, information theory, signal processing and linear algebra. This enabled the system to not only recognize the object, but also to localize where the object is present. The results discussed in this thesis evidently show that technical systems can be enhanced by following the biological paradigm.

KURZFASSUNG

Obwohl es für uns leicht erscheint, unsere Umwelt zu erkennen und mit ihr zu interagieren, fangen wir erst an zu verstehen, wie die visuellen Informationen im Gehirn in ihrem Ganzen verarbeitet werden. Die meisten der komplexen neurowissenschaftliche Modelle, beziehen ihre Daten aus empirisch gewonnenen psychologischen Studien und neurophysiologischer Forschung. Dabei haben diese Modelle das Ziel die unterschiedlichen Theorien zu evaluieren und verifizieren, ohne dabei eine potentielle praxisorientierte Anwendung zu betrachten. Rein technisch orientierte Bildsysteme sind hingegen begrenzt auf sehr spezifische Anwendungsgebiete wie Gesichtserkennung oder Objektverfolgung. Diese Systeme funktionieren jedoch nicht uneingeschränkt für jedes Szenario. Der Mensch hingegen ist in der Lage all diese technischen Systeme bei Weitem zu übertreffen.

Diese Dissertation verfolgt den Ansatz der effizienten Integration neurowissenschaftlicher Erkenntnisse in ein technisches Umfeld für die Verbesserung von Bildsystemen für zeitkritische Szenarien im Kontext von humanoider Robotik. Anstatt das Wissen über die Informationsverarbeitung im Gehirn zu replizieren, versucht die vorgestellte Arbeit jedoch den rein funktionellen Aspekt zu extrahieren, um technische Systeme zu verbessern.

Ein westenlicher wissenschaftlicher Beitrag dieser Dissertation ist die Analyse, Entwicklung und Evaluierung eines effizienten biologisch inspirierten Bildsystems, das in zeitkritischen Szenarien und Applikationen - wie in einem humanoiden Roboter - eingesetzt werden kann. Das vorgestellte Model beinhaltet drei fundamentale Gebiete des menschlichen Sehens: Visuelle Aufmerksamkeit (Visual Attention), objektbasierte Aufmerksamkeit (Object-based Attention), Objekterkennung (Object Recognition). Visuelle und objektbasierte Aufmerksamkeit spielen eine maßgebende Rolle in der Frage wie und was wir in unserem visuell wahrnehmen, indem sie die Fülle an vorhandenen Informationen selektieren und reduzieren.

Weitere wichtige Beiträge dieser Arbeit helfen das System zu realisieren: a) die Entwicklung eines neuen Visual Attention Systems, welches in der Lage ist andere, dem Stand der Technik entsprechende Systeme in Geschwindigkeit und Genauigkeit zu übertreffen; b) unsere Sampled Template Collation Methode zur effizienten Evaluierung verschiedener Bildregionen; c) ein neuer Object-based Attention Ansatz, welches in der Lage ist die Objekterkennung zu verbessern; d) unser Objecterkennungs Modell - eine Erweiterung des HMAX Modell, welches eine Abstraktion der neuronalen Informationsverarbeitung im visuellen Cortex darstellt. Das Modell wurde angepasst in Bezug auf Skalierbarkeit, Geschwindigkeit und Performanz, um es in zeitkritischen Szenarien nutzen zu können. Wir zeigen, dass durch die Integration neurowissenschaftlicher Erkenntnisse höhere Klassifikationsraten erzielt werden können; e) die Entwicklung einer Systemarchitektur für die Integration von Visual Attention, Object-based Attention und Objekterkennung, welche es dem humanoiden Roboter iCub ermöglicht auch sich schnell bewegende Objekte zu erkennen und segmentieren.

In Anlehnung an die hochparallele Informationsverarbeitung im Gehirn wurde das ganze System als parallel rechnendes und verteiltes System realisiert mit Unterstützung von Mehrkernprozessoren und Grafikprozessoren, um die vorhandene Rechenleistung effizient Nutzen zu können. Weitere Optimierungen beinhalten die Nutzung von Kenntnissen aus der Signalentdeckungstheorie, Informationstheorie, Signalverarbeitung und linearer Algebra. Das System ist zudem nicht nur fähig zu erkennen, was das Objekt ist, sonder auch, wo es ist. Die Ergebnisse dieser Dissertation zeigen, dass technische Systeme durch biologische inspirierte Ansätze verbessert werden können.

ACKNOWLEDGEMENTS

I would like to thank all the people that have been supporting me during the last years - without them it would have made my work impossible.

My special thanks go to my supervisor Gordon Cheng, who gave me the opportunity to carry out my research as a PhD student. He always encouraged me and provided a perfect environment in his institute for pursuing research. Special thanks go to Aleš Ude for overseeing my work and giving me the opportunity to visit the Jožef Stefan Institute and his group. I am also very grateful to all of my colleagues for their input and help. Without them work wouldn't have been so much fun.

I also want to thank my former diploma thesis supervisor Radu Rusu, who introduced me to the world of research and robotic vision and promoted me for pursuing a PhD. Many thanks also to my family for supporting me during my time as a student and later as a PhD student. A very special thank you to my loving wife, Karolina, for keeping my life in balance and supporting me without hesitation.

I gratefully acknowledge the funding sources that made my PhD possible: The DFG cluster of excellence Cognition for Technical Systems CoTeSys and BMBF through the Bernstein Center for Computational Neuroscience Munich BCCN.

Munich, 8th April 2015

Contents

PUBLICATIONS

- **Andreas Holzbach** and Gordon Cheng. *A Neuronal-inspired Computational Architecture for Spatio-Temporal Visual Processing.* Journal of Biological Cybernetics 108, p.249-259. June 2014.
- **Andreas Holzbach**, Gordon Cheng. *Object-based Attention using Efficient Sparse Sampled Template Collation.* Submitted to IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany. September 28 - October 3, 2015.
- **Andreas Holzbach**, Gordon Cheng. *A fast and scalable system for visual attention, object based attention and object recognition for humanoid robots.* IEEE-RAS International Conference on Humanoid Robots (Humanoids). Madrid, Spain. November 18-20, 2014.
- **Andreas Holzbach**, Gordon Cheng. *A scalable and efficient Method for Salient Region Detection Using Sampled Template Collation.* IEEE International Conference on Image Processing (ICIP). Paris, France. October 27-30, 2014.
- **Andreas Holzbach** and Gordon Cheng. *A concurrent real-time biologically-inspired visual object recognition system.* IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China. May 31 - June 7, 2014.
- Karinne Ramirez Amaro, Ewald Lutscher, **Andreas Holzbach** and Gordon Cheng. *iCub@ICS-TUM: Semantic Reasoning, Constrained Manipulation and Humanoid Vision enable on the iCub.* IEEE International Conference on Robotics and Automation (ICRA Workshop). Hong Kong, China. May 31 - June 7, 2014.
- **Andreas Holzbach**, Gordon Cheng. *Enhancing Object Recognition for Humanoid Robots through Time-Awareness.* IEEE-RAS International Conference on Humanoid Robots (Humanoids). Atlanta, USA. October 15-17, 2013.

- Zoltan Marton, Dejan Pangercic, Radu Rusu, **Andreas Holzbach**, & Michael Beetz. *Hierarchical object geometric categorization and appearance classification for mobile manipulation.* IEEE-RAS International Conference on Humanoid Robots (Humanoids). Nashville, USA. December 6-8, 2010.
- Radu Rusu, **Andreas Holzbach**, Michael Beetz, & Gary Bradski. *Detecting and segmenting objects for mobile manipulation.* IEEE International Conference on Computer Vision (ICCV Workshops). Kyoto, Japan. September 29 - October 2, 2009.
- Radu Rusu, **Andreas Holzbach**, Rosen Diankov, Gary Bradski, & Michael Beetz. *Perception for mobile manipulation and grasping using active stereo.* IEEE-RAS International Conference on Humanoid Robots (Humanoids). Paris, France. December 7-10, 2009.
- Radu Rusu, Zoltan Marton, Nico Blodow, **Andreas Holzbach**, & Michael Beetz. *Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments.* IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). October 10-15, 2009.
- Radu Rusu, **Andreas Holzbach**, Nico Blodow & Michael Beetz. *Fast Geometric Point Labeling using Conditional Random Fields.* IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). October 10-15, 2009.
- **Andreas Holzbach** and Gordon Cheng. *Sampled Template Collation for Simulating Visual Attention and Object-Based Selection.* Bernstein Conference. Göttingen, Germany. September 2-5, 2014.
- **Andreas Holzbach** and Gordon Cheng. *A Biologically-Motivated Approach to Online Learning of a Dictionary of Features from Natural Images for Computational Object Recognition.* Bernstein Conference. Tübingen, Germany. September 24-27, 2013.
- **Andreas Holzbach** and Gordon Cheng. *A Neurologically Motivated Computational Architecture for Real-Time Object Recognition.* Frontiers in Computational Neuroscience. 258, Bernstein Conference. München, Germany. September 12-14, 2012.

- **Andreas Holzbach** and Gordon Cheng. *An Information Theoretic Approach to an Entropy-Adaptive Neurobiologically Inspired Object Recognition Model.* Frontiers in Computational Neuroscience. 135, Bernstein Conference. Freiburg, Germany. October 4-6, 2011.

List of Figures

List of Tables

List of Algorithms

Chapter 1

INTRODUCTION

This chapter explains the motivation behind working on a biologically-inspired vision system and the problems that needed to be addressed and which were encountered. The chapter closes with the contribution of this work and an outline of this thesis.

1.1 Motivation

The idea of intelligent and human-like robots has always fascinated people. First introduced in science fiction literature and later movies, it now becomes more and more reality through ongoing research and development in the field of robotics. In industrial processes, robots have been integrated into the factory for years and are already vastly used for very specific automated tasks. The next big step will likely be robots for end users in home environments - as kitchen aid, to clean the house, to support elderly people in everyday tasks, to go shopping and maybe just as a companion. Robots can be built to easily extend our capabilities and to avoid the restriction of a biological body. They don't age, need no food, don't rely on air or water and are therefore perfectly suited especially for purposes in life-unfriendly environments like space exploration or rescue tasks for emergency situations in inhospitable terrain after an nuclear fallout, earthquake or fire.

Research could roughly be separated into two camps - empirical sciences like medicine, biology or chemistry and applied sciences like computer science and electrical or mechanical engineering. Former could be assigned to be more analytical research, especially in neurosciences where the structure and functionality of the human brain has been a focus of interest. Whereas latter sciences tend to be more on the engineering side where technical systems are built from scratch, focusing on very specific tasks. Recently researchers have drawn more attention to the enormous potential of neurobiological findings for the development of new technical models, which are able to exceed the capabilities of state-of-the-art systems.

The possibility to visually perceive the world is probably one of the most important abilities to any human being. As easy as it seems for us to recognize our surroundings and interact with our environment, we are still just at the beginning of understanding how the human vision system might really work. For a few decades, researchers have been able to investigate how visual information is processed in the retina and the visual cortex. Complex models have been built, drawing their foundation both from neurophysiology, as well as psychological empirical-obtained data. Despite ongoing analytical research, very few applicable models have shown to function in the real world. Most of the existing models aim at representing and evaluating different theorems, disregarding their real-world applicability - being in terms of efficiency, robustness and generality.

Pure technical-oriented vision systems however are confined to quite specific scenarios - like face recognition or tracking of objects - and seem to fail to function in a more general context while humans or animals still outperform any technical system by far. It is therefore worthwhile investigating the neuroscientific findings for its applicability as it could lead to a new generation of technical systems which can exceed the capabilities of state-of-the-art systems.

This work has been strongly motivated by the believe, that following the biological example might be the only possible way in succeeding to build intelligent robots with the capabilities of a human being.

1.2 Problem Description

One of the main challenges in pursuing a biologically-inspired approach is to solve the question of how to combine analytically obtained neurobiological models with technical systems in a way that the properties, the functionality and the advantages of the models are preserved. This is followed by the question if the particular model is suitable for a technical application in terms of performance, accuracy, and computational tractability.

If a model in its current form is not applicable for a technical utilization, is it possible to modify the model and how? In order to obtain the same functionality, is it necessary to replicate all the complexity we can find in the brain, the diversity of neurons and the way to process and transmit information? And if not, what level of abstraction can we apply to obtain and sustain the functionality while at the same time keeping the authenticity of the model to be a sound representation of parts of the brain.

Given that technical resources - like the resources in the brain - are limited, this can become a significant problem in the development of real-world systems, like humanoid robots.

Processing and interpreting huge amounts of data is a challenging problem - both from a computational and from an algorithm design perspective. Especially in the field of computer vision a large number of information carried in the pixels has to be processed in a short period of time before new information is available.

Each pixel might carry necessary and important data valuable for interpreting the current scene. The question is how to distinguish between redundant and important data and to extract only the necessary information in order to reduce the computational cost while at the same time being able to generate features which are unique, representative and invariant to different factors like lightning, orientation or scale.

This work is intended to contribute to the development of a scalable, modular and efficient vision system by using a biologically-inspired approach.

1.3 Contribution

This thesis presents an approach towards the efficient integration of neuroscientific knowledge into a technical environment for improving vision models in time-crucial real-world scenarios in the context of humanoid robotics. Instead of copying and computationally replicating what we know about the information processing in the brain, this thesis tries to grasp the pure functional aspect of it, in order to achieve similar results.

The main contribution of this thesis is the analysis, the development and the evaluation of an efficient biologically-inspired vision system which supplies a humanoid robot with the ability to visually perceive and understand its environment in an efficient and scalable manner. The proposed model integrates three essential parts of human vision: Visual attention, object-based attention and object recognition. Visual and object-based attention play a major role in how and what we perceive in our field of vision by selecting and reducing the available information. Both of which are essential for a fast and reactive vision system.

Six major contributions of this thesis helped to build this model:

1. The design of a new visual attention system, which outperforms state-of-the-art system in terms of accuracy, speed and complexity realized by
2. Our Sampled Template Collation method for efficiently evaluating different image regions, which is able to adapt to computational needs;
3. A new object-based attention system, which enhances object recognition;

4. Our object recognition model - an enhancement of a computational model called HMAX, which is an abstraction of the neural information processing in the visual cortex. The model was modified in terms of speed and performance in order to put it in a more technical context. We quantitatively and qualitatively show that by the integration of neuroscientific knowledge about neural information processing in the brain like lateral-inhibition and avoidance of entropic redundancy results in a higher classification accuracy and faster processing speed.

5. The development of a temporal reasoning framework which enables the system to classify over time and account for uncertainties in non-static real-world scenarios.

6. The system architecture for the efficient integration of visual attention, object-based attention and object recognition, which enabled the humanoid robot iCub to detect and segment even fast moving objects like a thrown ball.

1.4 Thesis Outline

The thesis consists seven chapters. It is thematically subdivided into chapters, which match the functional processing of the developed system (see figure 1.1):

Chapter 2. Related work.

This chapter, an overview on the related work is given. The chapter is split into three parts. (1) Visual Attention, (2) Object-Based Attention and (3) Object Recognition. Each of these chapters are divided into two subsections - the first gives a brief overview over the biological foundation of the specific topic, the second presents computational models that try to capture its functionality and behavior. Those models vary in their focus on biological accuracy and plausibility - mostly anticipated by life sciences - and technical applicability - more endorsed by engineering and computer sciences.

Chapter 3. Sampled Template Collation For Fast Saliency Maps Generation.

In this chapter, our visual attention system is presented. It's main advantages are a low computational complexity, online scalability and a high accuracy in predicting the human gaze which can compete with state-of-the-art models. First the core functionality of the system - Sampled Template Collation - is explained. The second section presents the experimental results of the system in regard to accuracy in predicting the fixation of human subjects and in regard to computational efficiency and scalability.

Chapter 4. Object-Based Attention using Sampled Template Collation.
Object-based attention explains the behavior of neural responses when an object is fixated. Visual stimuli are adjusted in favor of the particular object, which enhances the processing of the object's features. So far, only little research has been conducted towards the application of object-based attention in technical applications. In this chapter, our object-based attention system is presented. The method we developed is based on our sampled template collation model. The advantages are the low computational complexity and the improved object recognition results due to the segmentation. The first section describes how sampled template collation is used for object segregation. The second section presents the visual segmentation results and the improved object classification performance.

Chapter 5. Enhancing a Computational Model for Object RecognitionModHMAX.

This chapter will introduce the object recognition system developed in this thesis. It consists of ModHMAX, a for time-crucial applications enhanced modification of HMAX and the concept of temporal reasoning, which introduces time to static recognition models and presents a more realistic approach to biologically-inspired object recognition. The chapter starts with a discussion about the use of 3D information in object recognition.

Chapter 6. A System Architecture for Visual Attention, Object Segregation and Object Recognition.
This chapter proposes an architecture for the integration of visual attention, object-based attention and object recognition for active camera systems. We describe how the single modules are integrated into a software framework and how the communication and information processing is handled between the modules. We successfully enable the humanoid robot iCub to adjust the gaze to the most salient point using out visual attention system based on sampled template collation. The fixated object is then fed to our object-based attention system for object segregation. The segmented object is then classified using our ModHMAX approach.

Chapter 7. Conclusions.
Finally, we present a summary of this thesis as well as some further discussion of

our system and contributions. We also give an outlook and suggestions of possible future work and improvements.

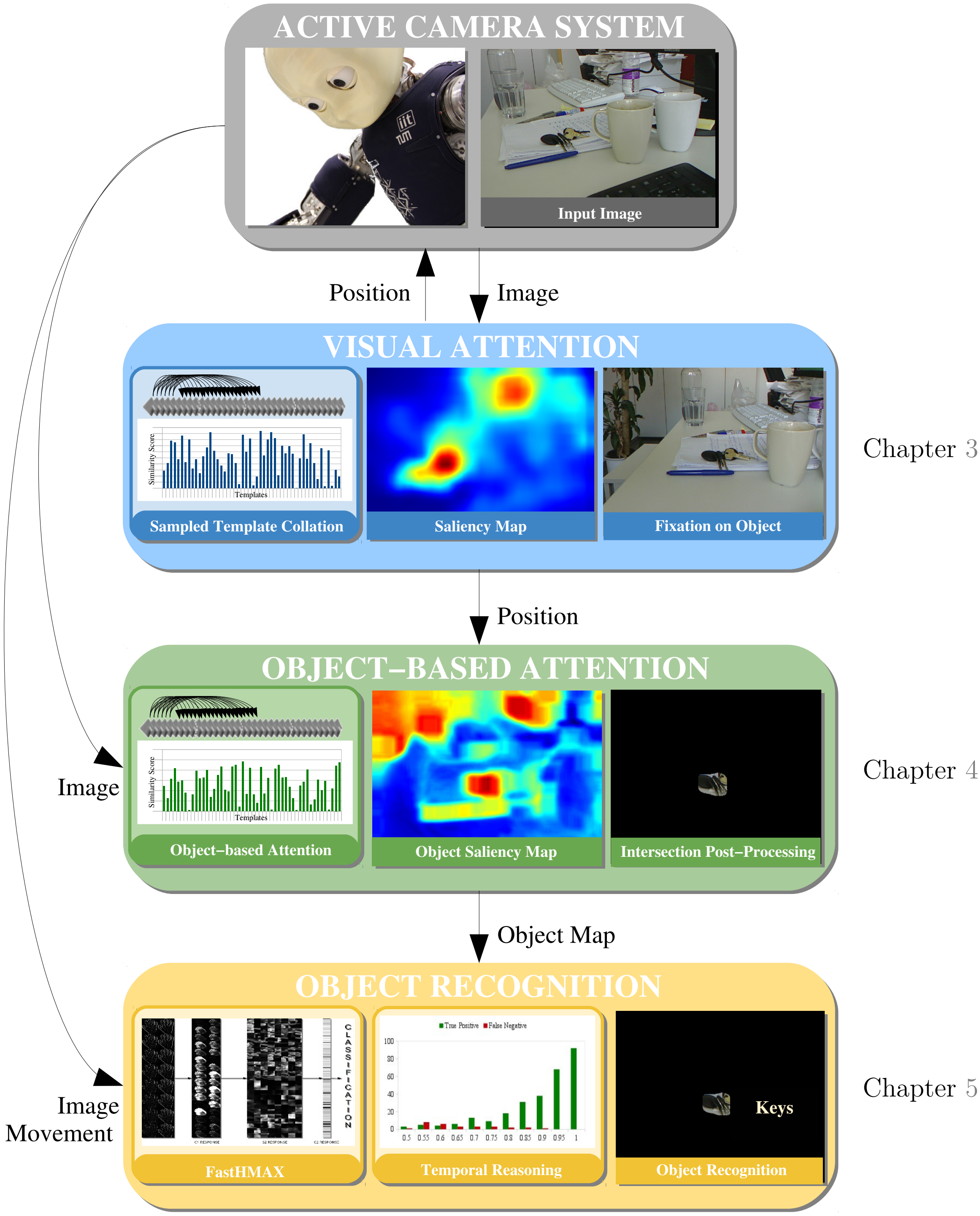


Figure 1.1 Thesis Outline. The visual attention system is described in chapter 3, the object-based attention system in chapter 4 and the object recognition system in 5

Chapter 2

RELATED WORK

In this chapter, an overview on the related work is given. The chapter is split into three parts. (1) Visual Attention, (2) Object-Based Attention and (3) Object Recognition. Each of these chapters are divided into two subsections - the first gives a brief overview over the biological foundation of the specific topic, the second presents computational models that try to capture its functionality and behavior. Those models vary in their focus on biological accuracy and plausibility - mostly anticipated by life sciences - and technical applicability - more endorsed by engineering and computer sciences.

The most challenging problem in robotics is arguably vision, and we stand to gain a great deal in following the biological system as a guideline for the design of artificial visual system for robots, especially humanoid robots [Ude and Cheng, 2004; Ude et al., 2005; Goerick et al., 2005; Ude et al., 2008b]. Humans are capable of detecting and recognizing objects under the most complex circumstances. They can easily identify objects under most lighting conditions, different orientations, shapes or sizes [Duhamel et al., 1997; Anderson et al., 2000; Booth and Rolls, 1998]. Even objects in clutter pose little problems, in contrast to state-of-the-art computer-based object recognition systems , which struggle to perform adequately under varying situations [Riesenhuber and Poggio, 1999; Serre et al., 2007b]. Therefore, it only makes sense – and perhaps is the only successful way – to analyze how the visual system in biological systems works and use that knowledge for modeling those mechanisms to build a more likely effective and robust object recognition system.

The human brain contains more than 10 billion neurons and more than 10 trillion synapses, making up networks and subnetworks of immense complexity [Yantis, 2008]. Recently, due to a deeper understanding of information processing in the brain and due to more powerful computational resources, the vision and robotics community started building more and more systems which gain their inspiration and functionality from biological models. Be it for building robots by studying the human corpus [Pfeifer et al., 2007], or for building robotic insects [Wood, 2008], or to integrate intelligence into humanoid robots [Bar-Cohen and Breazeal, 2003]. Or to enhance common techniques like face recognition by using biologically-inspired features [Meyers and Wolf, 2007]. The widely applied SIFT features [Lowe, 1999] are also inspired by neurons in the inferior temporal cortex. Some research draw more attention to active-vision systems, which have been used to solve different vision problems like: object recognition [Chen et al., 2011; Bevec and Ude, 2012; Andreopoulos et al., 2011; Goerick et al., 2005]; visual search [Rasolzadeh et al., 2010; Halverson and Hornof, 2012]; visual attention [Siagian and Itti, 2007]; or visual tracking [Mahadevan and Vasconcelos, 2013]. It has also been investigated how to integrate object recognition [Ude et al., 2008b, 2004] and visual attention also with a focus on the aspect of computational complexity [Ude et al., 2005].

2.1 Visual Attention

The amount of sensory information provided by the visual world is immense and computationally expensive to process in it's whole. It is therefore useful and important to filter out currently unnecessary information to avoid an overload of sensory data and to reduce the computational payload. Visual attention is a concept that origins from life sciences, especially psychology, where researchers try to explain the unconscious selection process of visual information in humans. The nervous system emphasizes information which seem to be more important or more interesting compared to other stimuli and which is subsequently passed on for a more detailed investigation. This concept of information reduction and detecting regions of interest or uniqueness is obviously very useful for technical applications and has been exploited in a vast variety of different vision tasks like feature detection, image segmentation [Mishra et al., 2009a,b], image matching [Siagian and Itti, 2007], image and video compression [Cheng et al., 2011], object detection [Goferman et al., 2012] or tracking [Frintrop, 2010]. In the first subsection the biological background is briefly discussed, followed by the second subsection which presents some state-of-the-art computational models for visual attention.

2.1.1 Biological Background

The research on visual attention[1] is arguably one of the most broadly studied topics in a wide area of research fields from psychology, cognitive neuroscience or even computer science. Visual attention describes the unconscious selection of a subset of all visually observed information. The perceived stimuli are biased in favor of the most salient ones, which usually is a small set of stimuli that show different patterns in comparison to the rest of the stimuli.

Visual Attention has been under intense investigation in research probably because it is present in all everyday situations involving vision. A bird in the sky draws our attention towards it, because the color sticks out from the blue background. So does a red apple on a tree, because the green of the leaves surround it. Traffic signs are designed to stick

[1] In literature visual attention is used to describe spatial attention - in contrast to object-based attention, although visual attention refers to the whole process of attention in the visual cortex. In this work visual and spatial attention mean the same.

out from the surroundings in order to draw a driver's attention to it. Marketing agency have been using concepts from visual attention to create appealing advertisements.

Carrasco investigated the visual attention related publications of the last 25 years and found an almost exponential increase in the number of articles in scientific journals (see figure 2.1) [Carrasco, 2011]. Borji and Itti show in [Borji and Itti, 2013] a taxonomy of different attentional studies (see Figure 2.2).

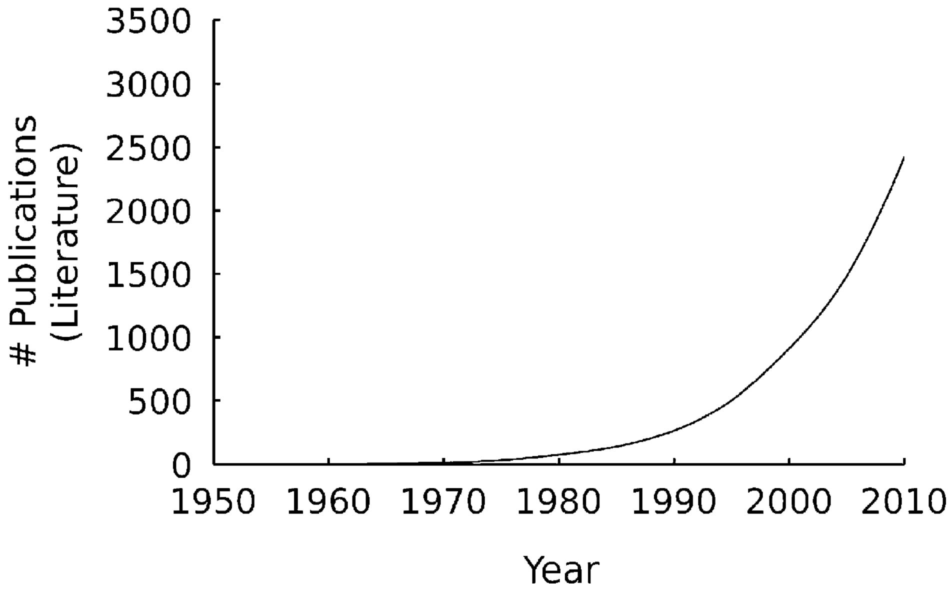


Figure 2.1 Number of articles on visual attention published in all scientific journals [Carrasco, 2011].

Mark et al. [Mark et al., 2007] describe attention as the state of selectively processing simultaneous sources of information. In the context on visual attention, this means, that attention enables us to concentrate on one object over others in our visual field.

There are two camps that grant different rolls to attentional processing [Kanwisher and Wojciulik, 2000]. Recent research believes that preattentive vision perceptually analyses the entire scene to a higher level, even including object identification. A subset of this information is then selected for further analysis and response planning. The traditional research on the other hand, considers the preattentional process as a very basic mechanism without any involvement in higher cognitive functions. They believe that object recognition is only possible with focused attention, after preattentive processing.

This work will focus on the classical view of visual attention, because of its long history, broader acceptance and vast variety of models which have been developed to describe

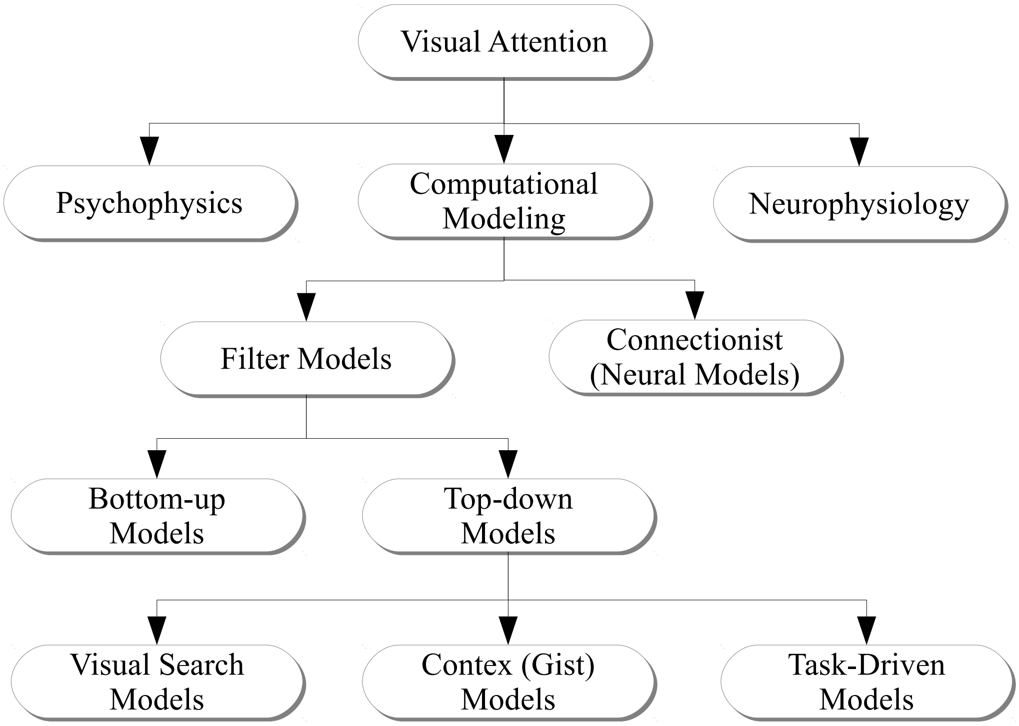


Figure 2.2 Taxonomy of visual attention studies [Borji and Itti, 2013].

it. Although those models don't integrate a higher preattentive analysis like suggested in recent work, the prediction rate of those models comes very close to the human's performance [Judd et al., 2012], which suggests the correctness of the classical approach.

Two approaches have been intensely discussed in attentional research, namely bottom-up and top-down [Connor et al., 2004]. Bottom-up mechanisms operate on the raw sensory input without any conscious shift of attention to salient areas in the visual field. These low level features can be e.g. differences in color, motion, orientation, lightness. Top-down mechanism on the other hand involve higher cognitive strategies, biasing the raw input toward specific features e.g if we are looking for our keys, or searching for a specific tool in the kitchen. Top-down mechanisms are also responsible for the strong attentional bias towards faces, humans or animals. This phenomenon supports the idea of some kind of preattentive object identification.

One of the first and probably most prominent concepts of bottom-up visual attention is Treisman's Feature Integration Theory [Treisman and Gelade, 1980]. In contrast to preattentive object identification theories, Treisman suggests that features are registered early, automatically, and in parallel during the preattentive stage, while objects are identified separately and at later step in processing called focused attention stage. The proposed features include separable dimensions like shape and color and local elements or parts like lines, edges or curves which are separately analyzed and the integrated into a complex whole representation. The visual scene is accordingly coded into separable dimensions, such as color, orientation, spatial frequency, brightness and direction of movement. The features that are present at the fixated area of attention are combined to form an object. Figure 2.3 gives an example of the different feature maps involved in the visual attention process.

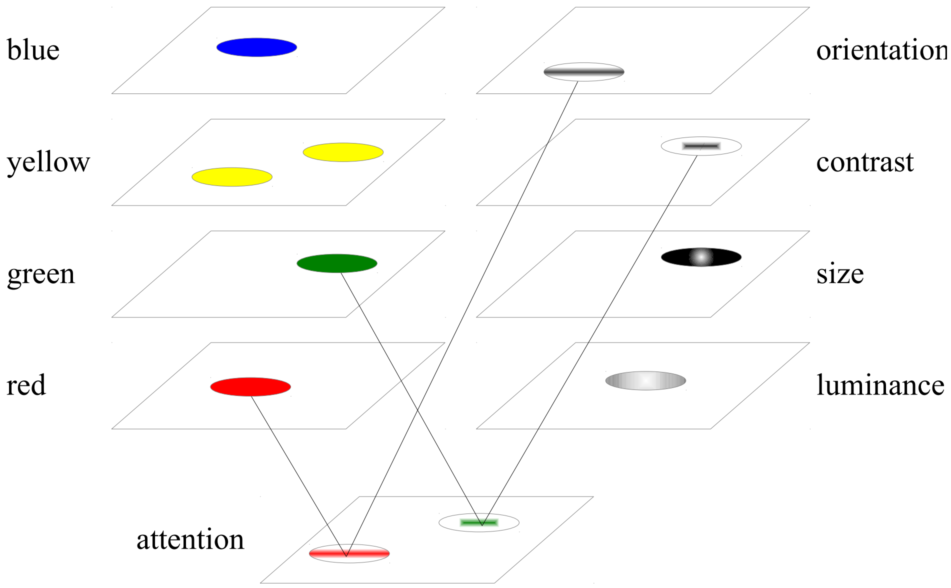


Figure 2.3 Feature Integration Theory example with individual feature maps: The basic colors blue, yellow, green, red and orientation, contrast, size and luminance.

More recently visual attention has been investigated on a neuroscientific level using functional magnetic resonance imaging (fMRI). Kanwisher and Wojciulik show in [Kanwisher and Wojciulik, 2000] where visual attention happens in the brain using functional neuroimaging which provides new and more detailed insights in the neural processing. They found out that attention not only modulates the gain on incoming visual information like suggested by bottom-up models, more importantly attention can also add a pure top-down signal that increases baseline activity in striate and extrastriate cortex. Attention can also select locations, features or objects under different conditions.

They conclude that attention affects the processing already at the first stage of cortical information processing in the primary visual cortex.

Itti and Koch give a simplified overview over the involved brain areas during visual attention [Itti and Koch, 2001] (See figure 2.4). The information is processed along two parallel and hierarchical streams. The dorsal stream (including the posterior parietal cortex) is responsible for the spatial localization of the visual information and therefore for directing the attention and gaze towards objects of interest. The ventral stream (including the inferotemporal cortex) on the other hand is responsible for the recognition and identification of objects.

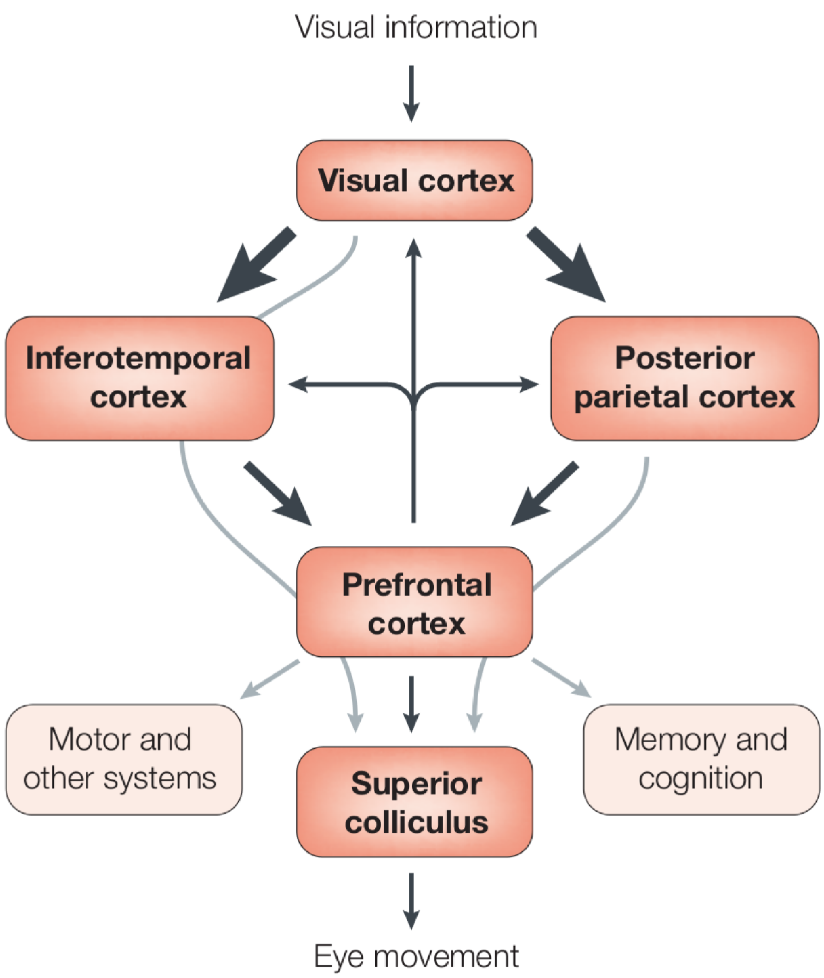


Figure 2.4 Neural Mechanism and brain regions involved in visual attention processing and motor control of the eye (Image taken from [Itti and Koch, 2001]).

2.1.2 Visual Attention Models

There exist a vast number of visual attention models. The first models that were created can be classified as cognitive models, that focus on biological plausibility and have their roots in Treisman's Feature Integration Theory. Borji and Itti [Borji and Itti, 2013] grouped the most prominent state-of-the-art models into different categories: Information Theoretic Models, Graphical Models, Spectral Analysis Models, Pattern Classification Models, Bayesian Models or Decision Theoretic Models. In this subsection one of the probably best known models by Itty and Koch [Itti et al., 1998] will be presented as well as Harel's, Koch's, and Perona's Graph-based Visual Saliency model (GBVS) [Harel et al., 2006]. A distributed visual attention model for humanoid robots by Ude et al. [Ude et al., 2005] concludes this section to give an example of a technical application of visual attention and its realization in real world scenarios.

Itti and Koch present in [Itti et al., 1998] a saliency-based visual attention model for rapid scene analysis (see figure 2.5), based on the Feature Integration Theory. It decomposes the input image into multiscale image pyramids generated with a low-pass Gaussian filter. The pyramids have different modalities: color, intensity and orientation - latter created using edge detecting Gabor filters at different orientations. Then the center-surround differences are calculated. These operations are similar to the functionality of visual receptive fields. The neurons are most sensitive in the center, while the surrounding can inhibit the response. This behavior particularly contributes to the detection of locations that differ from their surrounding area. All resulting feature maps are then normalized and combined to a final saliency map. The model was designed under the premise of biological plausibility in accordance with the anatomy of the visual system in macaque monkeys.

The Graph-based visual saliency model (GBVS) was proposed by Harel et al. [Harel et al., 2006]. They argue that their Markovian graph-based approach is inspired by the neural communication in the visual cortex. The architecture consists of two stages: In the first step activation maps of certain feature channels are created, and are then normalized in the second step, in a way which highlights conspicuity and combination with other maps. The feature maps are created from basic feature channels of the image, like color, orientation, intensity. The activation map is created calculating dissimilarities between pixels in those feature maps using

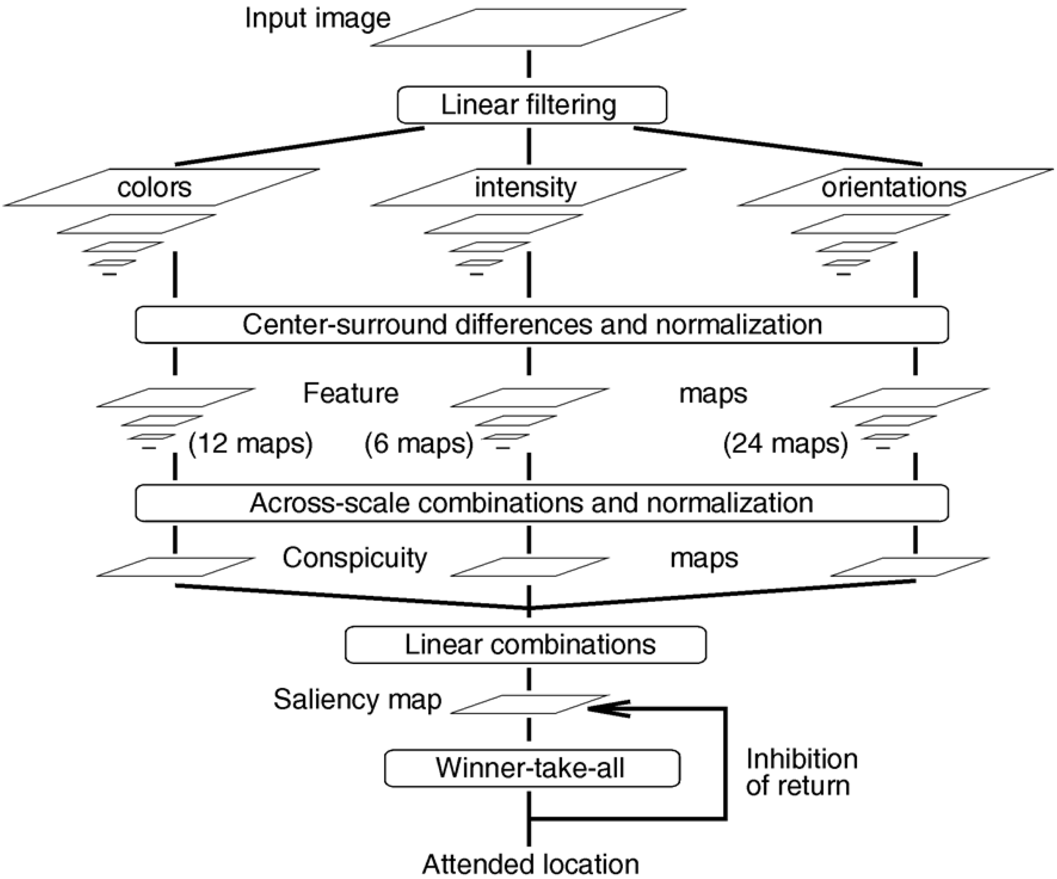


Figure 2.5 General architecture of Itti and Koch's model. (Image taken from [Itti et al., 1998]).

$$d((x_1, y_1), (x_2, y_2)) = \left| log \frac{M(x_1, y_1)}{M(x_2, y_2)} \right| \tag{2.1}$$

In the next step a fully-connected, directed graph is built by connecting the point of interest with all other points in M. The edges of the graph G are weighted using the feature dissimilarity score and a Gaussian-like distance weight F:

$$\begin{aligned} &w((x_1, y_1), (x_2, y_2)) = d((x_1, y_1), (x_2, y_2)) * F(x_1 - x_2, y_1 - y_2), \\ &with \quad F(a, b) = exp(-\frac{a^2 + b^2}{2\sigma^2}) \end{aligned} \tag{2.2}$$

Now a Markov chain is defined on the graph G by normalizing the weight of all edges of every node to 1. Each node can now be regarded as a Markovian state and its edges as transition probabilities of the Markov chain. The higher the weight, the higher is the transition probability, meaning it is more likely to move on to a state, that was more different and accordingly more unique in the feature map. The equilibrium distribution of this chain would accumulate mass at nodes with high dissimilarities, because of the higher transition probabilities. The resulting mass distribution creates the activation map A. The second part of the approach aims on promoting areas with a high amount of dissimilarities by concentrating the masses from the activation map A. This step is similar to the previous one. A Graph is created using subsets of pixels in A as nodes. The edge weights are calculated using equation 2.2. Harel et al. argue, that this method seems to behave better compared to a Differences of Gaussians or nonlinear interactions approach. The normalization step can be carried out multiple times to account for local maxima. The final saliency map is now created by summing the normalized single channel activation maps.

Many of the approaches to visual saliency are computational expensive and complex (see comparison in [Cheng et al., 2011]), making those models less suitable in real-world scenarios. More recent systems, however, are more focused on computational performance, like [Lin Zhang, 2013] or Cheng et al.'s work in [Cheng et al., 2011]. Latter compare different models by their computation times for building a saliency map and propose a fast model of their own based on regional contrast. The most related model in regard to our template sampling approach is Erdem and Erdem's work in [Erdem and Erdem, 2013]. They compare covariances of non-overlapping neighbored image regions to compute the saliency map.

Ude et al. applied a visual attention system on a humanoid robot [Ude et al., 2005] and showed how computational expensive models can be implemented in a parallel fashion to achieve low-latent real time performance. Their visual attention system is built on Itti and Koch's architecture. They apply the same feature cues intensity, orientation, and color but extended the model with motion and disparity. Two very important features to detect salient areas in a real world scenario. The human brain is strongly biased towards motion, which could be explained by evolutionary processes, where it was important to detect a moving enemy or prey in order to survive. By introducing disparity to the system they account for the fact, that close objects or objects in reach are more interesting or important that distant objects. To handle the computational cost, they split up the cues on a cluster of PCs and combine the resulting maps later on another PC. Ude et. al use synchronization schemes to handle the possibility of delay in the image processing, otherwise unsynchronized images could be combined which would result in erroneous saliency maps. They applied their system on a humanoid robot with an active eye system by driving the gaze towards the most salient point generated by their system.

2.2 Object-Based Attention

Object-based attention tries to explain what happens in the brain after a saccade, when an object is fixated. In this step the focused object or area seems to stick out suppressing the surrounding, even in cluttered scenes. The attended object is consciously perceived, which is also described as selective attention in contrast to previous unconscious perception.

It is obvious that this step contributes towards object recognition by separating the attended object from the surrounding and biasing features in favor of the attended objects. It can therefore be regarded as some kind of segmentational process, similar to segmentation in computer vision.

In this section the biological background of object-based attention is presented. Object-based attention systems for technical applications has however hardly been investigated, here we present work by Walther et al., whose research has focused particularly on that topic.

2.2.1 Biological Background

Because of the long history of research on spatial attention, theories on object-based attention are not as mature and represent just a small fraction of attentional research. Recent research, however, has demonstrated the importance of objects in organizing (or segregating) visual scenes, guiding attentional selection and for object recognition [Vecera, 2000; Walther and Koch, 2006].

Desimone and Duncan describe two basic phenomena that define the problem of visual attention [Desimone and Duncan, 1995]. The first one is the limited capacity for processing the information available on the retina. The second one is the ability to filter out currently unnecessary information, which enhances the visual representation of objects, even if spatially occluded in cluttered real-world scenarios.

Yantis [Yantis, 2008] describes two categories that influence visual selection. One is bottom-up, involuntary or unconsciously, and stimulus-driven and depends on primitive physical appearance of an object. The second kind of influence on selection is top-down and depends on a more cognitive decision. He emphasizes that selective attention is required when the visual system is confronted with typically cluttered natural scenes.

Cohen and Tong [Cohen and Tong, 2013] describe object-based attention as a pattern-specific attentional filtering in the visual cortex, meaning that patterns which don't belong to the attented object - like color or texture - are suppressed. The neural activity patterns in early visual areas are strongly biased in favor of the attended object.

Together with foveal vision and the much higher density of cones in the center of the fovea, the phenomenon of object-based attention contributes towards the recognition of objects in higher cortical areas [Walther et al., 2005], as it helps to segment the object from irrelevant clutter [Walther and Koch, 2006].

Ungerleider et al [Ungerleider and G, 2000] also state that in everyday life, the scenes we view are typically cluttered with many different objects, but the capacity of the visual system to process information about multiple objects is limited. They explain that the different objects in the visual field compete for neural presentation, because of the limited processing capacity in the visual cortex. This competition is biased by bottom-up mechanisms driven by primitive features and top-down influences, such as selective attention.

Kastner et al. [Kastner and Ungerleider, 2001] explain this behavior more detailed. At the neural level, competition among multiple stimuli is evidenced by the mutual suppression of their visually evoked responses and occurs most strongly at the level of the receptive field. Functional brain imaging studies reveal that biasing signals due to selective attention can modulate neural activity in visual cortex not only in the presence, but also in the absence of visual stimulation. Subjects that are shown two different objects and that are asked to identify two dissimilar attributes at the same time (e.g., color of one and orientation of the other) perform worse than if the task had been performed with only a single object. They conclude that multiple objects present at the same time in the visual field compete for neural representation due to limited processing resources. Sensory suppression among multiple stimuli present at the same time in the visual field has been found in several areas of the visual cortex, including areas V2, V4, the middle temporal (MT) and medial superior temporal (MST) areas, and inferior temporal (IT) cortex.

Similar statements can also be found by Mangun in [Mangun, 1995]. He states that visual selective attention improves our perception and performance by modifying sensory inputs at an relatively early stage of processing. These spatial filters alter the inputs to higher stages of visual analysis that are responsible for feature extraction and ultimately object perception and recognition, and thus provide physiological evidence for early precategorical selection during visual attention.

Desimone advocates the Biased Competition Theory [Desimone, 1998] which suggests that the visual processing in the brain can be biased by other mental processes such as Bottom-up or Top-down. He proposes a model which is comprised of five main tenets:

1. Objects in the visual field compete for cell responses in the visual cortex. If, for example, two stimuli are presented simultaneously within the visual field, both neural representations are initially activated in parallel. Neural responses in that region will be determined by a competitive interaction between those stimuli. These interactions will be mutually suppressive on average.

2. Competitive interactions are strongest in a given cortical area when competing stimuli activate cells in the same local region of cortex. Receptive fields which receive two competing stimuli will react the strongest.

3. Competitive interactions can be biased in favor of one stimulus in a cluttered field by virtue of many different mechanisms, rather than by a single overall attentional control system. These mechanisms are bottom-up (e.g. one stimulus has greater novelty or has a higher contrast than another) as well as top-down driven.

4. The feedback bias is not purely spatial and can be driven by stimuli possessing a specific relevant feature, like color, texture, contrast or shape. This non-spatial driven feedback refers to behavioral tasks like visual search, where the sum of specific features are biased over location. When a subject is e.g. asked to look for a red object in his visual field, stimuli which react to red are emphasized and stimuli of other colors suppressed.

5. The main source of the top-down biasing inputs to ventral stream areas in extrastriate cortex derives from structures involved in working memory, specifically prefrontal cortex.

Proulx et al. [Proulx and Egeth, 2008] describes the Biased Competition Theory as a competition of objects for cortical representation in a mutually inhibitory network, which is biased in favor of the attended item. They also mention that bottom-up and top-down information are the two sources of information which allow an object to be processed over other objects.

The Biased Competition Theory has broadly found acceptance in related research communities until today. More recent research like Vecera's paper on Object-based Segregation [Vecera, 2000] also state that the biased competition model has been useful in describing a range of behavioral and neurobiological data from visual search experiments that rely on spatial attention, which suggests that the general approach of combining stimulus information and goal-related information may provide an accurate description of many attentional phenomena.

Vecera's work is focused on object-based segregation, which refers to the visual process responsible for determining which visual features combine to form shapes which contains the object and separates it from its surroundings. Object segregation is synonymous with perceptual organization, the term used in conjunction with the gestalt principles of visual organization [Koffka, 2013]. Object segregation and object-based attention are interrelated. Before a shape can be selected, the features of the shape first must be segregated from features of other shapes to some extent. The ability to perform figure-ground segregation and distinguish foreground shapes from background regions also involves object segregation processes.

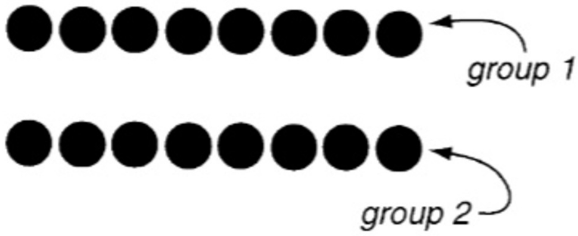
In Figure 2.6 an example of figure-ground segregation is given. Object-based attention could be directed to either the black region or the white region. Attentional selection is more efficient if attention is directed to a single region than to multiple regions. Because any visual scene contains many objects that compete with one another as they are being segregated and compete for attention, the visual system must be capable of allocating processing to one object or region over others. This allocation is achieved by biasing processing toward one object or region. These biases provide a resolution for the competition between objects or regions; this competition between objects occurs within both segregation processes and object attention processes. For example, the two regions

Figure 2.6 An example of figure-ground segregation in object-based attention. The black region appears to be more salient than the white region or the gray background [Vecera, 2000].

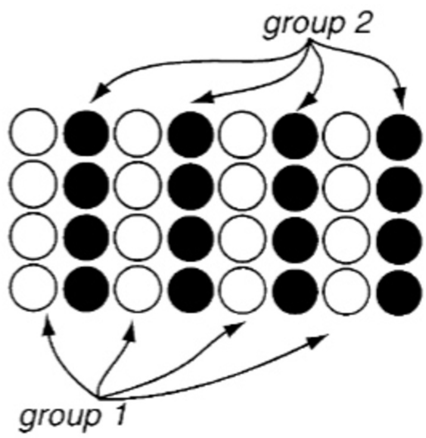
in Figure 2.6 compete with one another for neural representation. Observers tend to perceive the symmetric black region as figure because symmetry acts as a bottom-up bias in the figure-ground competition and favors symmetric regions as figure (and asymmetric regions as ground). The biased competition account attempts to explain (1) how some objects or regions become more salient figures or perceptual groups and (2) how some objects are selected over others.

Vecera [Vecera, 2000] mentions, that two sources of competition occur in natural multi-object scenes. First is a competition within object-based segregation processes and second a competition within object-based attentional processes. The outcome of the first competition is a perceptual group that is more salient than others; the outcome of the second is the actual selection process of the perceptual group. These two sources of competition are obviously highly interrelated, they are however discussed as separate in the visual perception literature. In this work, there is no distinction made between those two terms.

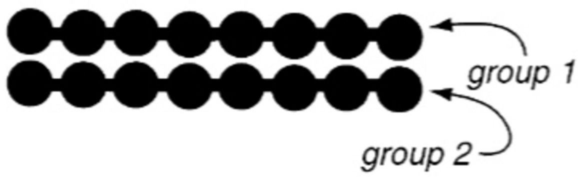
An important source of bottom-up information influencing object segregation are so called image cues, which allow the visual system to determine if two features origin from the same object or different object. The gestalt principles of organization, shown in figure ??, are bottom-up cues that are useful for visual segregation



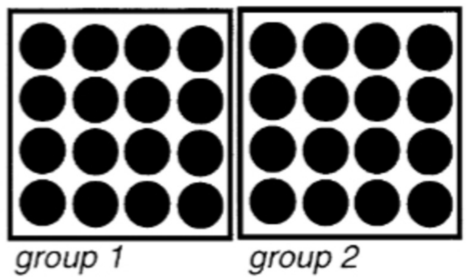
(a) Proximity



(b) Similarity



(c) Connectedness



(d) Common Region

Figure 2.7 Some of the gestalt cues of perceptual organization. These cues provide bottom-up information helping the visual system to segregate salient objects from their surrounding. (a) The proximity cue allows close features to group to higher order information (The circles form horizontal rows of elements.) (b) The similarity cue groups features on the basis of similar primitives (Circles with same luminance are grouped together.) (c) The connectedness cue groups features that are physically connected to one another. (d) The common region cue groups features that are within a shared region (Images taken from [Vecera, 2000]).

The gestalt principles are a basic model of visual perception, because it does not integrate any concept of learning, which is evidently a essential part of neural processing. Through experience the visual system learns representations of natural scenes. Features of the same color, for example green of leaves, are more likely to belong together as features which are unknown to the perceptual system. Similar features are assigned to the same object and features which are different are expected to belong to a different object or shape.

Object-based segregation and object recognition are obviously strongly interrelated. Without object segregation a multi-object scene, which basically represents all natural scenes, would be much harder to classify, if not entirely impossible. Petersen et al. [Peterson, 1994] assume an even stronger interrelation between both processes, that is based on feed-back information processing. They believe that object recognition partly happens prior to figure-ground segregation. Objects are preliminarily identified by a prefigural recognition process and then again influence figure-ground segregation. This prefigural assumption stands in contrast to most hierarchical accounts of perceptual segregation and object recognition that place segregation processes prior to recognition processes like Marr [Marr, 1982] or Biederman [Biederman, 1987].

2.2.2 Object-Based Attention Models

Sun and Fisher present an object-based visual attention model for computer vision [Sun and Fisher, 2003] that extends Duncan's Integrated Competition Hypothesis. In contrast to existing feature-driven models, they extend their system to be object-driven. They describe two new mechanisms in their proposed uniform framework: The first computes the visual salience of objects and groupings; the second implements hierarchical selectivity of attentional shift. They state three theoretical aspects found in modern literature, which are brought together in their model: 1.) Integrated competition for visual attention. 2.) Bottom-up and top-down interaction and 3.) Hierarchical selectivity of visual attention.

Walther et al present in [Walther et al., 2005] a selective visual attention system and address the problem of object learning in cluttered scenes. They propose a method for the selection of salient regions which are likely to contain objects, based on bottom-up visual attention. They apply the method on unsupervised one-shot learning of single objects in clutter and show that it can strongly improve learning and recognition performance. Their attention system is based on Itti et al's approach explained earlier, where the result of the Winner-Take-All operation is used for object segmentation and then further processed with Lowe's SIFT features. They validated their approach and achieved a false positive rate of 0.8% compared to 6.8% for random patch selection.

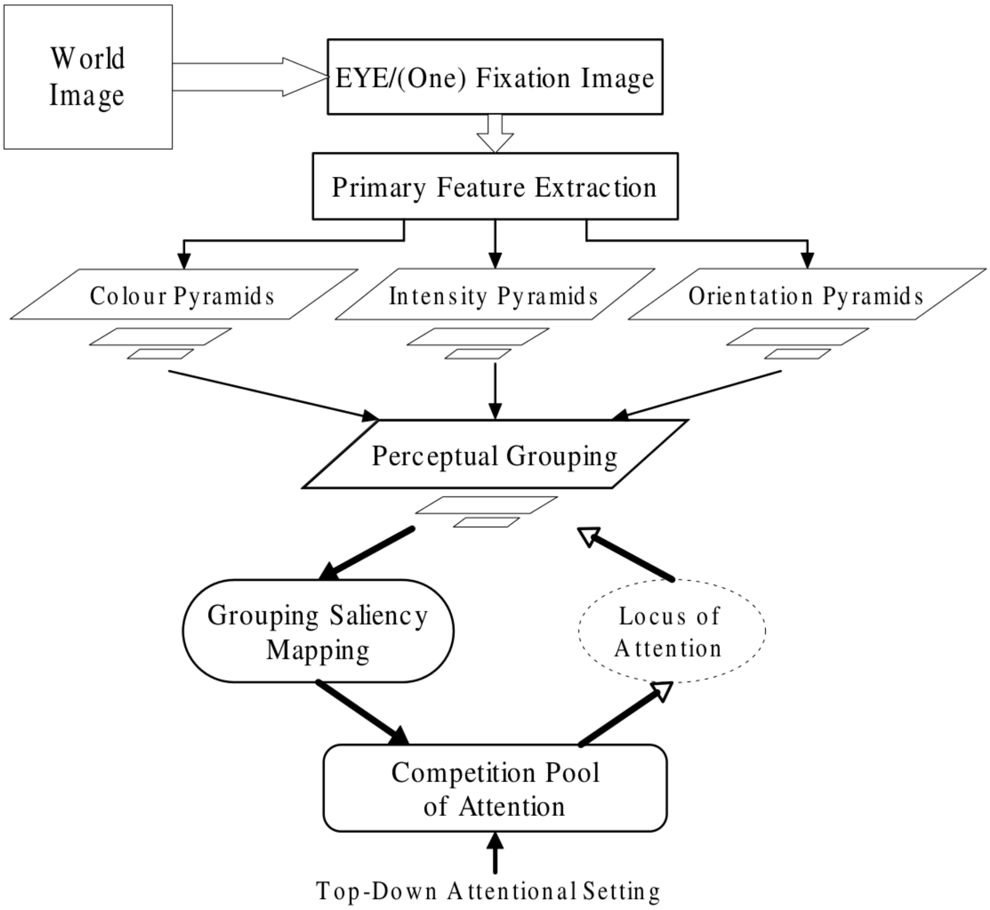


Figure 2.8 Sun and Fisher's model of object-based attention, integrating both top-down and bottom-up features. (Image taken from [Sun and Fisher, 2003].)

2.3 Object Recognition

In the last couple of years there has been an increase in biologically-inspired hierarchical models for object recognition, due to a deeper understanding of information processing in the brain [Thomure et al., 2010; Serre et al., 2007b; Poggio et al., 2011]. Some of these models have also been applied to enhance common techniques like face recognition by using biologically-inspired features [Meyers and Wolf, 2007]. Some research draw more attention to active-vision systems, which have been used to solve different vision problems like: object recognition [Chen et al., 2011; Andreopoulos et al., 2011; Goerick et al., 2005; Wersing and Körner, 2003]; visual search [Rasolzadeh et al., 2010; Halverson and Hornof, 2012]; visual attention [Siagian and Itti, 2007]; or visual tracking [Mahadevan and Vasconcelos, 2013]. It has also been investigated how to integrate object recognition

[Ude et al., 2008a, 2004] and visual attention also with a focus on the aspect of computational complexity [Ude et al., 2005]. Especially the HMAX model [Riesenhuber and Poggio, 1999] has been investigated and modified in multiple publications [Serre et al., 2007b; Moreno and Mar, 2007; Mutch and Lowe, 2006; Theriault et al., 2011].

2.3.1 Biological Background

As easy as it seems for mammals to process visual information and infer their environment in less than 200 milliseconds [Thorpe et al., 1996; Serre et al., 2007b; Tovee, 1994; Reinagel and Reid, 2000], despite a extensive detailed knowledge about the anatomical architecture of the cortical areas and its functional organization [Van Essen et al., 1992], we still have much to learn about the brain's vision system. For a few decades, researchers have been able to investigate how visual information is processed in the retina and the visual cortex [Nassi and Callaway, 2009; Vinje and Gallant, 2000; Huth et al., 2012]. Complex models have been built, drawing their foundation from both neurophysiology and psychological, empirically obtained data [Serre et al., 2007a; Riesenhuber and Poggio, 1999, 2000; Gustavo Deco, 2004].

The hierarchical structure of the visual cortex and the interconnections between the individual visual areas are highly distributed and parallel. Figure 2.9 visualizes the immense complexity, which makes it difficult to understand the full functionality of the visual cortex, although the identification of connections between the cortical areas and the emphasis of the hierarchical structure have been well examined [Van Essen, D.C. et al., 2001; Felleman, D.J. and Van Essen, 1991]. DeYoe and Van Essen point out the manifold relationships between sensory cues and perceived attributes [DeYoe and Van Essen, 1988]. Nassi and Callaway provide an overview over recent research about distributed parallel processing strategies in the visual system [Nassi and Callaway, 2009]. Sharpee et al. demonstrate in [Sharpee et al., 2004] a model which maximizes the mutual information between the neural responses. They applied their approach to responses of simple and complex cells in the visual cortex and obtained realistic estimates of the relevant dimensions by maximizing information.

DiCarlo et al. [DiCarlo et al., 2012] suggest, that the brain is able to rapidly recognize objects through a cascade of reflexive, largely feedforward computations that builds a powerful neural representation in the inferior temporal cortex. Figure 2.10 visualizes a simplified model of cortical areas involved in vision processing. The number of neurons in each cortical area decreases along the pathway from 190 million in V1 to 150 million in V2 and 65M in V4. In the last layer the inferior temporal cortex receives about 10 million neural responses for each representation. It takes about 100ms for the visual data to arrive at the inferior temporal cortex, which is consistent with rapid scene analysis studies, where subjects are asked to identify scene which are shown for a short period of time.

2.3.2 Object Recognition Models

There exists a large number of different object recognition models or models for feature generation. In this section three state-of-the-art models are presented. The first and most prominent one is arguably the Lowe's scale-invariant feature transform (SIFT). The second one are deep convolutional neural networks inspired by LeCun's work, which have recently shown to outperform any existing system by far. The third one is the convolutional neural network HMAX by Poggio, Riesenhuber and Serre, which is inspired by the information processing in the visual cortex.

All three models are more or less biologically-inspired to some degree. It is however important to mention, that the neural basis of those models is very tenuous from a neuroscientific point of view. Those models were created not with a focus on biological plausibility - the neuroscientific model is rather stripped down to some functional aspects of the processing.

SIFT features for example adapt the shift-invariance of the complex neurons in the visual cortex. Deep Convolutional Neural Networks (DCNN) adapt the hierarchical structure in the brain and the learning of neural activity towards specific patterns. HMAX in contrast has a lower hierarchy with four layers and fixed filter set for convolution. It already performs well with a low number of training images whereas DCNN needs much more input images to train their weights and filters. A lack of this huge amount of data was responsible for a minor performance in these DCNNs. Only recently DCNNs could show their enormous capabilities with the development of large image databases like ImageNet[1] and the introduction of Rectified Linear Units as activation functions to the architecture [Glorot et al., 2011]. The training however takes a very long time, with about one week on a GPU cluster for training about 19 million parameters.

[1] http://www.image-net.org/

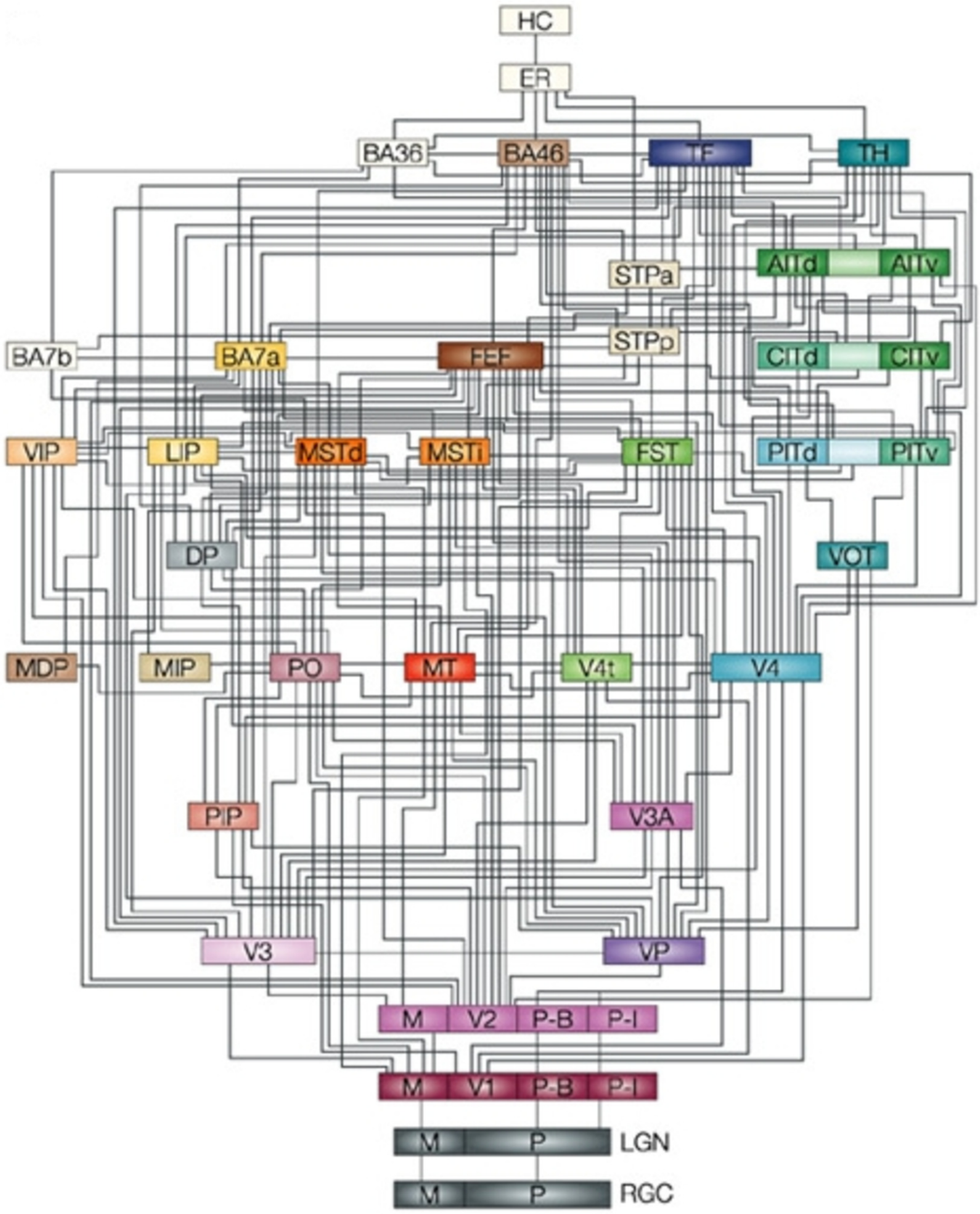


Figure 2.9 Hierarchical representation of the visual areas in the brain. Felleman and van Essen distinguish 32 visual cortical areas connected by 187 linkages. [Felleman, D.J. and Van Essen, 1991].

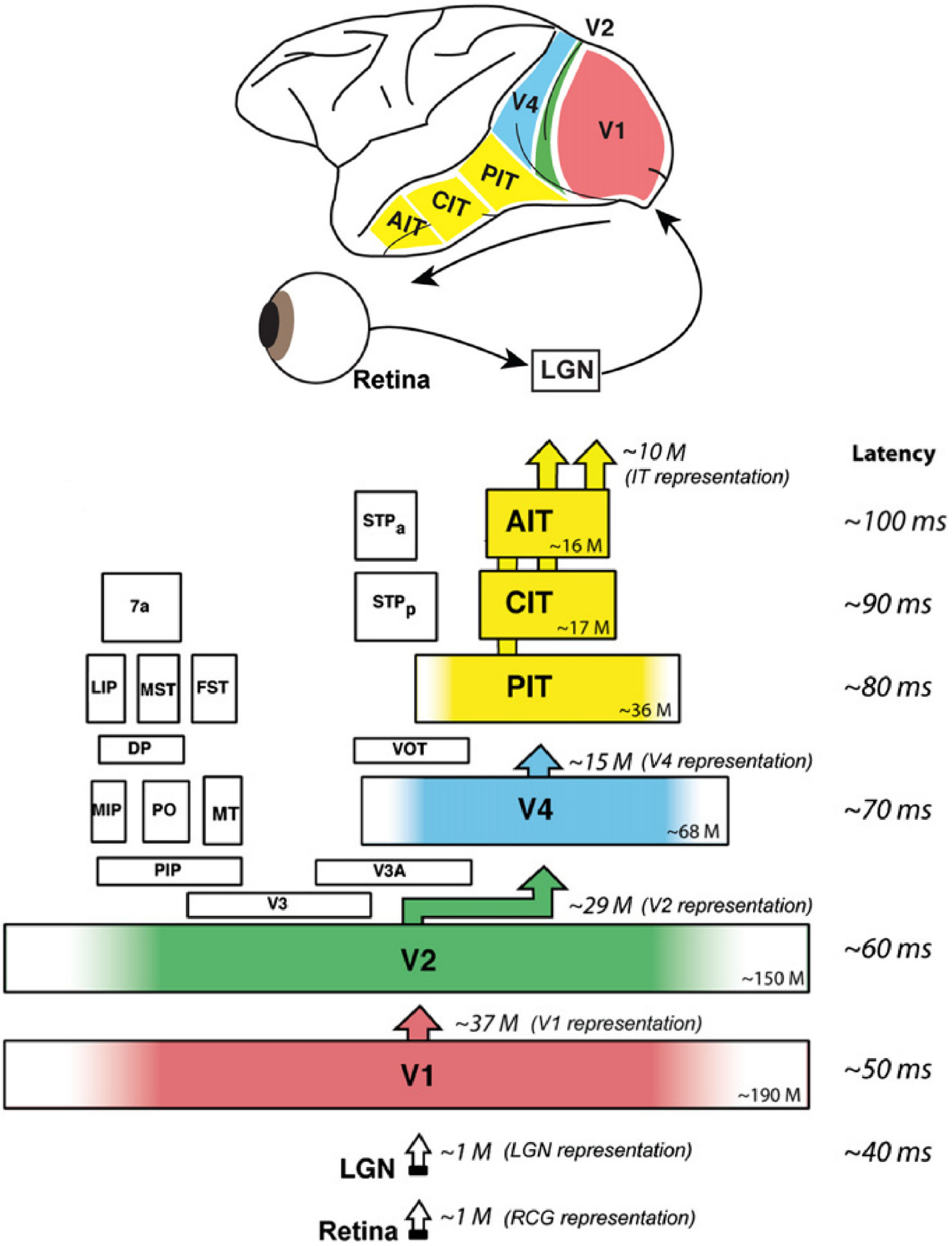


Figure 2.10 Cortical areas involved in vision processing in the brain. The top image shows the locations in a macaque's cortical area that show reaction during an object recognition process. The visual information origins from the retina and is processed through the ventral stream to LGN, V1, V2, V4, PIT, CIT and finally AIT. The bottom image displays the latency and the processing direction in the ventral stream. The size of the different cortical areas are proportional to the rectangles. The approximate number of neurons is shown in the right corner of each rectangle. The approximate number of neurons involved in each representation is shown above the rectangles. (Images from [DiCarlo et al., 2012]).

Table 2.1 Classification accuracy averaged over 15 classes. 150 training examples per class [Ciliberto et al., 2013]. SIFT, Bag-of-Words and Sparse Coding.

	k-NN(%)	RLS (%)	SVM (%)
SIFT	39.9	-	-
BOW	60.6	84.7	83.6
SC	68.2	87.7	86.6
HMAX	80.7	86.5	89.1

Our work on the object recognition part is strongly based on HMAX, as it allows

1. for a quick training with little training data,
2. the computational advantages for training and processing, which is a crucial requirement for robotics in real life scenarios,
3. its possibilities for improvements and modifications and
4. it proofed superior to state-of-the-art object recognition systems like SIFT.

Ciliberto et al. [Ciliberto et al., 2013] compared HMAX to three commonly used object recognition approaches: SIFT, Bag-of-Words and Sparse Coding. They evaluated the different algorithms using a image database of 15 classes. HMAX outperforms the other approaches is most cases (see table 2.1). Serre et al. [Serre et al., 2007b] compared HMAX features to SIFT for different number of training example and features. They also experienced a significant difference between the two approaches (see 2.11). Moreno et al. [Moreno et al., 2007] performed another comparative study between SIFT and HMAX. They evaluated various variants of SIFT and HMAX: Original HMAX, HMAX sampled at DoG, SIFT non-rotation-invariant, original SIFT, SIFT-Gabor and SIFT-Gabor non-rotation-invariant.They conclude that HMAX performs better than any of the SIFT variants in all of their different experiments.

2.3.2.1 Scale-Invariant Feature Transform (SIFT)

The SIFT algorithm was developed by D.Lowe [Lowe, 1999] and creates features that are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. The algorithm can be described in four steps.

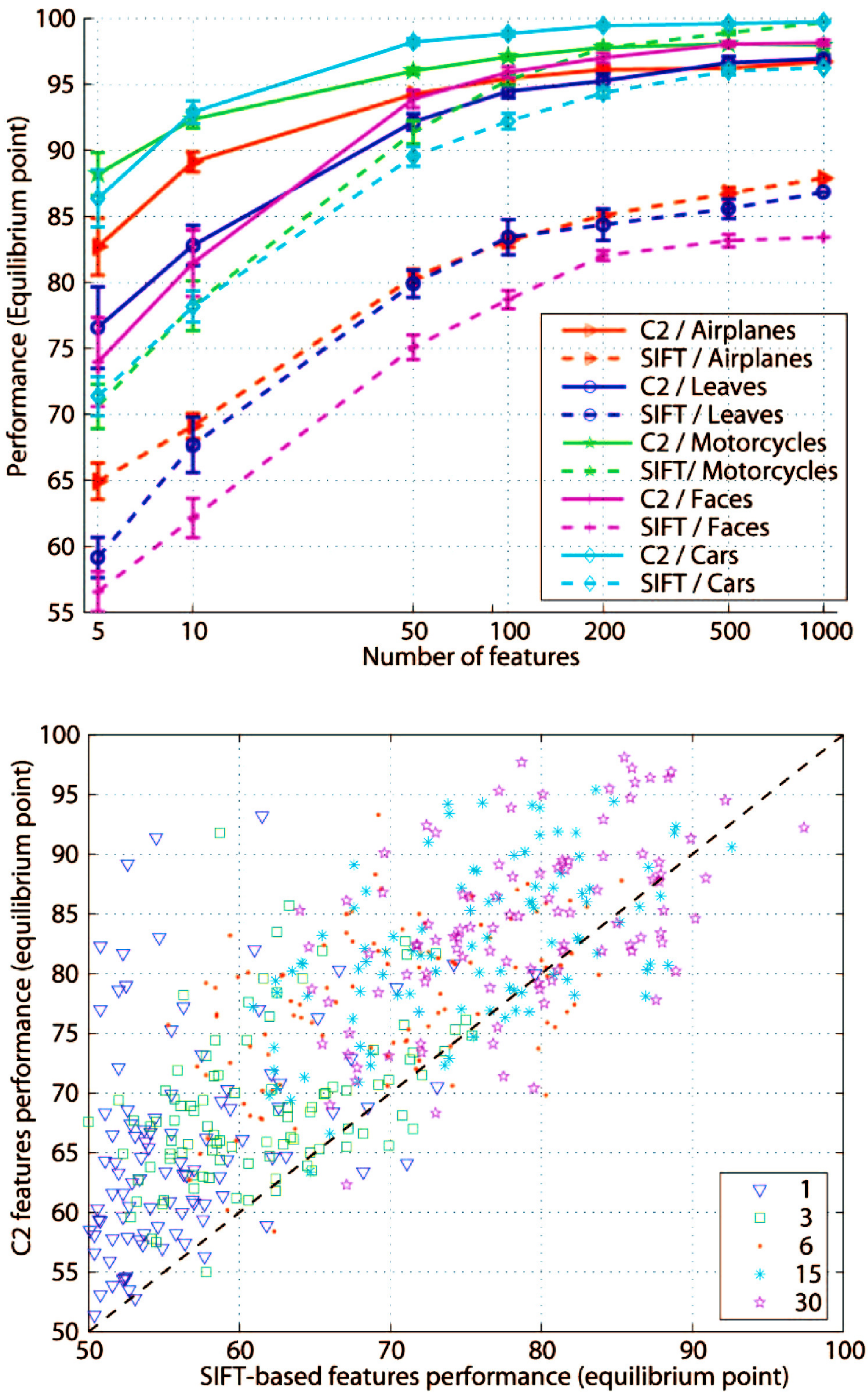


Figure 2.11 Comparison between SIFT and HMAX C_2 features for different number of features (top) and different number of training examples (bottom). (Plot from [Serre et al., 2007b])

.

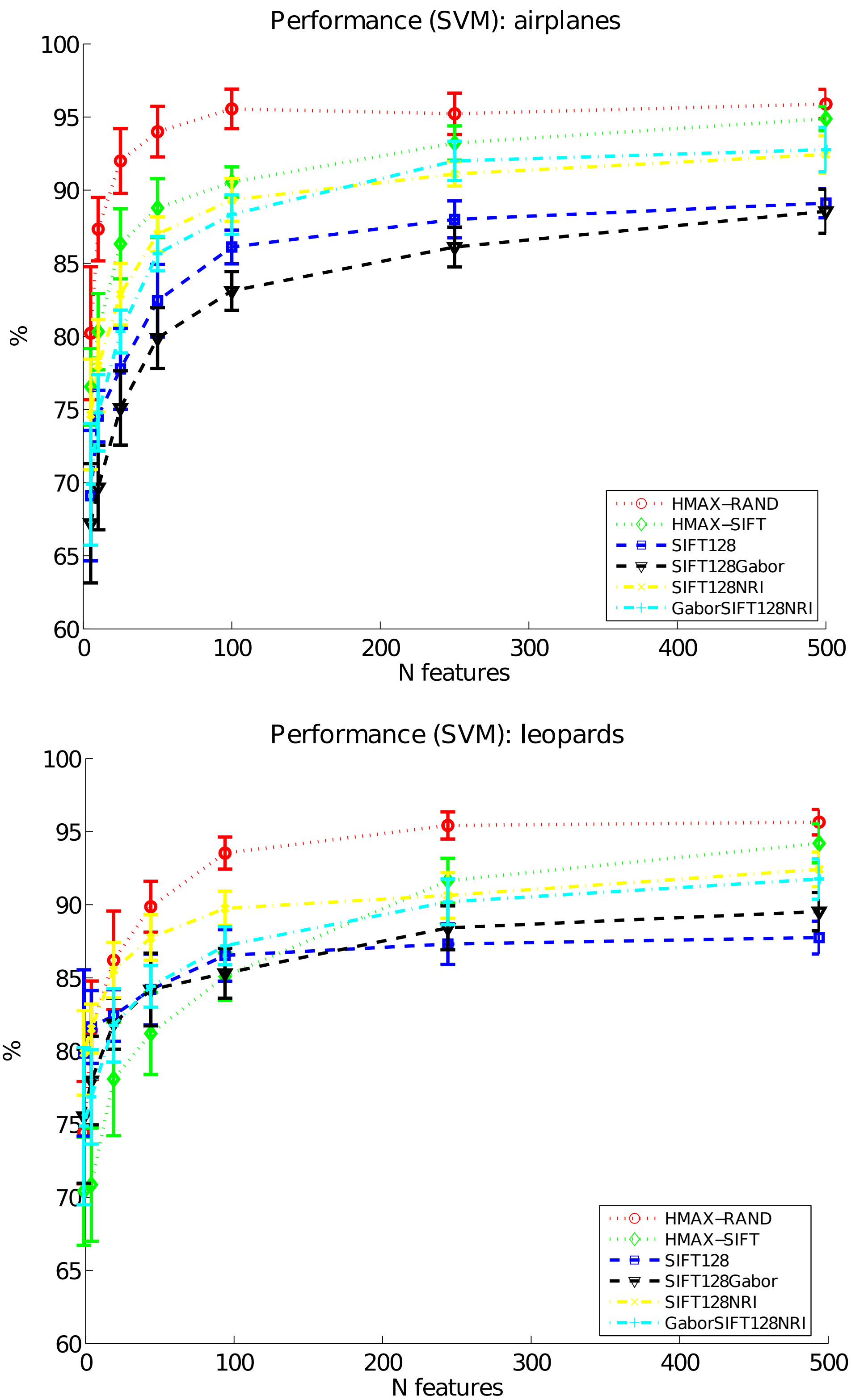


Figure 2.12 Comparison between SIFT and HMAX for object detection performance of airplanes(top) and leopards(bottom) depending on the number of features (Plot from [Moreno et al., 2007]).

.

First, the scale-space extrema are detected using Difference of Gaussians, which creates a set of keypoints with (x, y, σ), with σ being the variance of the applied Gaussian filter. In the second step the keypoints are localized with sub-pixel accuracy using Taylor series expansion of scale space. Points with intensities lower than a chosen threshold are rejected, which eliminates keypoint candidates with a low contrast. To achieve invariance to rotation the third step involves an orientation assignment to each keypoint. Therefore an orientation histogram is created with the gradient magnitude of the neighborhood around the keypoint. Any value above 80% is used to calculate the orientation. In the forth step the keypoint descriptor is created using a 16x4x4 neighborhood with 8 bin orientation histograms resulting in a total of 128 bin values. The keypoint descriptor is the vector of this histogram. To remove false-positive keypoint descriptor matches between two images, the two nearest neighbor matches are compared. If the match is greater than 80%, they are rejected. This eliminates around 90% of false matches while only 5% of true positives are removed.

2.3.2.2 Deep Convolutional Neural Networks

In the last couple years Deep Convolutional Neural Networks (DCNN) showed remarkable performance on recognizing thousands of different object categories using large image databases for training. LeCun [LeCun et al., 1998] describes that DCCN are built from three architectural ideas: local receptive fields, shared weights and spatial sub-sampling.

The local receptive fields refer to convolution operations with an edge-like filter. This filtering is similar to the reaction of simple cells to edges in their receptive field in the visual cortex. The operation which defines the spatial sub-sampling is max pooling. It is similar to complex cell responses, which are local-spatially invariant to particular responses. A DCNN's architecture consists of multiple alternating layers of convolution and max pooling with different filters and weights, which are learned during training stage. In 2012 Krizhevsky et al. presented a DCCN which exceeded any previous object recognition systems by far (15.3% error rate compared to the second best with 26.2%, [Krizhevsky et al., 2012]). Figure 2.13 shows their DCCN architecture. Their network has 60 million parameters and 650,000 neurons with 5 convolutional and max pooling layers, followed by three fully-connected layers.

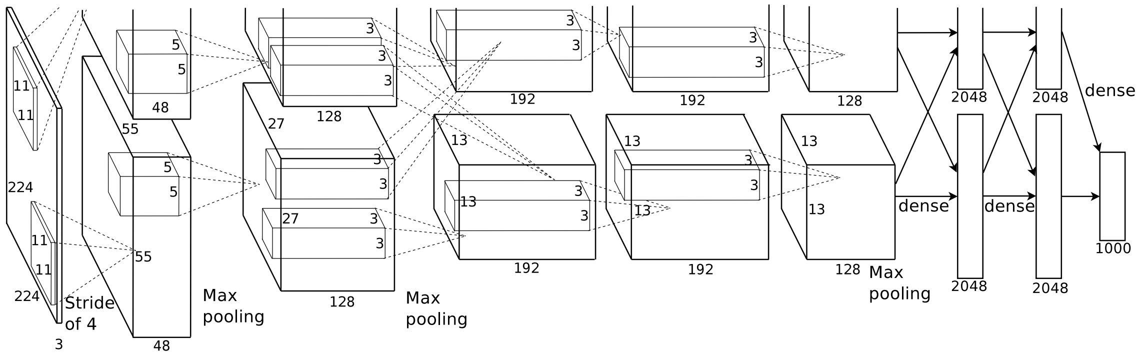


Figure 2.13 Krizhevsky et al.'s architecture of a convolutional neural network. (Images from [Krizhevsky et al., 2012])

.

2.3.2.3 The HMAX Model

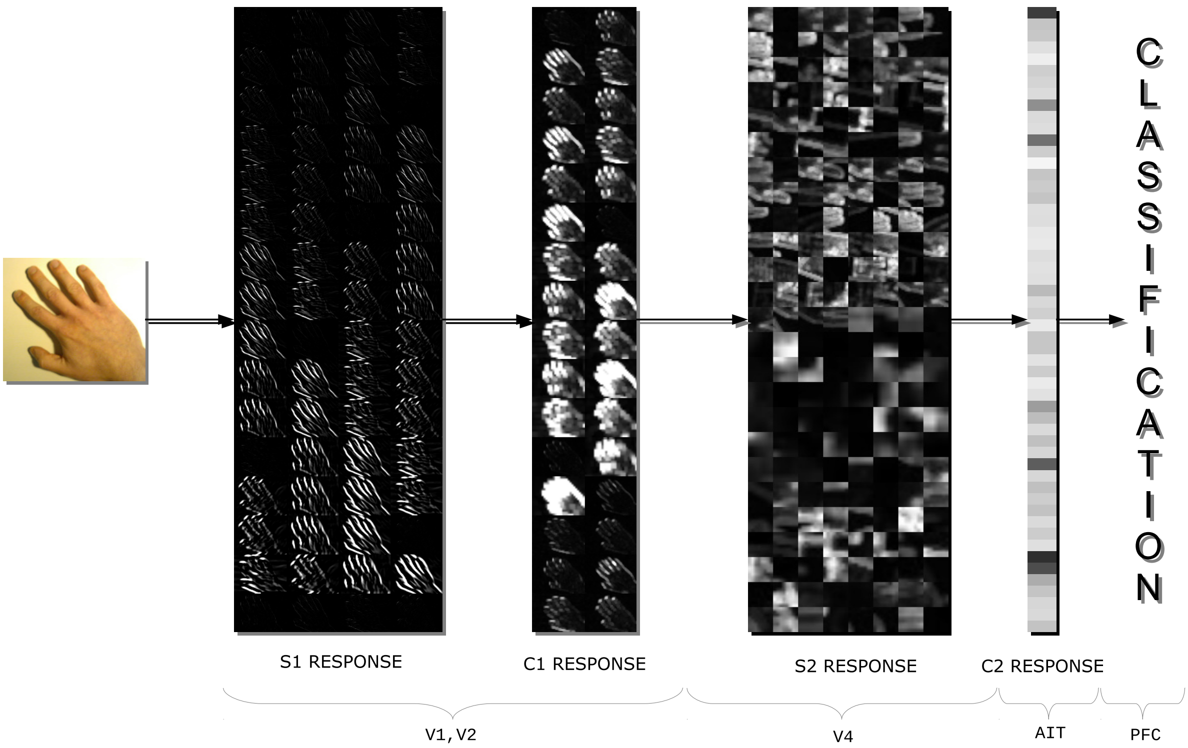


Figure 2.14 Functional Overview of the hierarchical object recognition architecture HMAX. The single layers process the information in parallel and pass it on to the next layer.

The HMAX model by Riesenhuber and Poggio [Riesenhuber and Poggio, 1999; Serre et al., 2007b] matches closely with empirically-obtained data from psychological [Hubel and Wiesel, 1968] and neurobiological studies [Tanigawa et al., 2010]. It models the ventral pathway in the areas of the visual cortex V1, V2 and V4 and its hierarchical feed-forward structure for visual object recognition. This strict feed-forward processing is a simplification to the actual classification process in the visual cortex where feedback plays an important role. Riesenhuber and Poggio model the behavior of simple and complex cells found by [Hubel and Wiesel, 1968] in the visual cortex in four alternating layers of simple cells (S1, S2) and complex cells (C1, C2). Figure 2.14 visualizes the different layers and outputs. Here we briefly illustrate the standard HMAX model presented in [Serre et al., 2007b]

S1 Layer The first layer is based on a representation of simple cells in V1 which react to oriented edges and bars in their receptive field. The response of these cells is quite similar to Gabor filters with specific parameters according their tuning of orientation and frequency;.

A Gabor Filter Bank of 64 filters is used for convolution with the input image to create a representation of the S1 receptive field response. Serre et. al. [Serre et al., 2007b] apply parameters, which resemble the response of the actual V1 parafoveal simple cells in the visual cortex; corresponding to neurophysiological data in [De Valois et al., 1982].

C1 Layer Complex cells have a a larger receptive field than simple cells and add some degree of spatial invariance and shift tolerance to the system. They gain input from two S1 filter outputs of same scale band and same orientation. Their functionality can be described as a max pooling operation or a moving maximum over two filter outputs of S1; They keep only the maximum value of two neighbored (of same band) responses of the previous S1 layer within a sliding window.

S2 Layer In the third layer a set of templates is matched against the response from the previous layer C1. The templates are sampled over the whole receptive fields from a set of randomly chosen images. The layer models the composite feature cells in V4 [Riesenhuber et al., 1999].

C2 Layer Like in C1, the complex composite cells in the C2 layer perform a max operation over all the template responses across all scales. The operation removes all position and scale information resulting in global invariance. The whole response is a complex feature vector which can be used to train and test a classifier.

2.4 Summary

In this chapter we presented a brief introduction and overview of neuroscientific and technical research related to this thesis. The three main modules were visual attention, object-based attention and object recognition. All three are of strong scientific relevance, both in neurosciences like biology or psychology and in technical sciences like computer vision or engineering. First we gave an introduction in the biological foundation of visual attention and presented some computational models later. Object-based attention was explained in more detail, as it is less known and investigated in the research community. Finally object recognition models relevant to this work were introduced.

In the following chapters, our contribution to these three fields is presented. First we introduce Sampled Template Collation for visual attention, then object-based attention and finally object recognition.

Chapter 3

SAMPLED TEMPLATE COLLATION FOR FAST SALIENCY MAPS GENERATION

The majority of visual attention systems were created with a focus on biological plausibility and a high accuracy in predicting the human saccades. Only little attention has been drawn to efficiency and scalability. In this chapter, our visual attention system is presented. It's main advantages are a low computational complexity, online scalability and a high accuracy in predicting the human gaze which can compete with state-of-the-art models.

In the first section the core functionality of the system - Sampled Template Collation - is explained. The second section presents the experimental results of the system in regard to accuracy in predicting the fixation of human subjects and in regard to computational efficiency and scalability.

Salient region detection is a broadly investigated research area, because it is concerns a wide field of scientific disciplines. Life sciences like psychology [Lamberts and Goldstone, 2004] or neuroscience [Ungerleider and G, 2000] are interested in analyzing and predicting why salient regions are attractive to the human brain and how the neural processing is involved in this decision [Bundesen et al., 2005]. Numerous computational models have been proposed trying to model visual attention and predicting which areas will be favored over others. Those saliency estimators can loosely be separated into biologically based, computational, or a combination of both which builds the majority of models [Borji and Itti, 2013]. The nervous system emphasizes information which seem to be more important or more interesting compared to other stimuli and which is subsequently passed on for a more detailed investigation. This concept of information reduction and detecting regions of interest or uniqueness is obviously very useful for technical applications and has been exploited in a vast variety of different vision tasks like feature detection, image segmentation [Mishra et al., 2009a,b], image matching [Siagian and Itti, 2009], image and video compression [Guo and Zhang, 2010], object detection [Goferman et al., 2012] or tracking [Frintrop, 2010].

Our model calculates the saliency map by sampling templates randomly over the image. Each template is then compared to the other templates by calculating a dissimilarity score. Higher scores mean lower similarity, and vice versa lower responses higher similarity. Templates with a higher overall dissimilarity score therefore originate from areas in the image which stick out from the rest and are in some kind unique. We consider these areas salient and use the templates' dissimilarity score to generate our saliency maps. See figure 3.1 for an overview of the model.

3.1 Sampling

First we sample templates from random positions on the image. For the evaluation we used templates of three different sizes (8,16,24). The different sizes account for the different dimensions a salient region might have. In other systems this behavior is achieved using for example Difference of Gaussian pyramids or subsampling. Using only one single template size however, doesn't affect the AUC(area under curve) of the receiver operator characteristics (ROC) score. We experienced only about 0.02% difference in the AUC score when using only one size. The number of sampled templates can be adjusted according to computational or accuracy requirements. Less templates can be

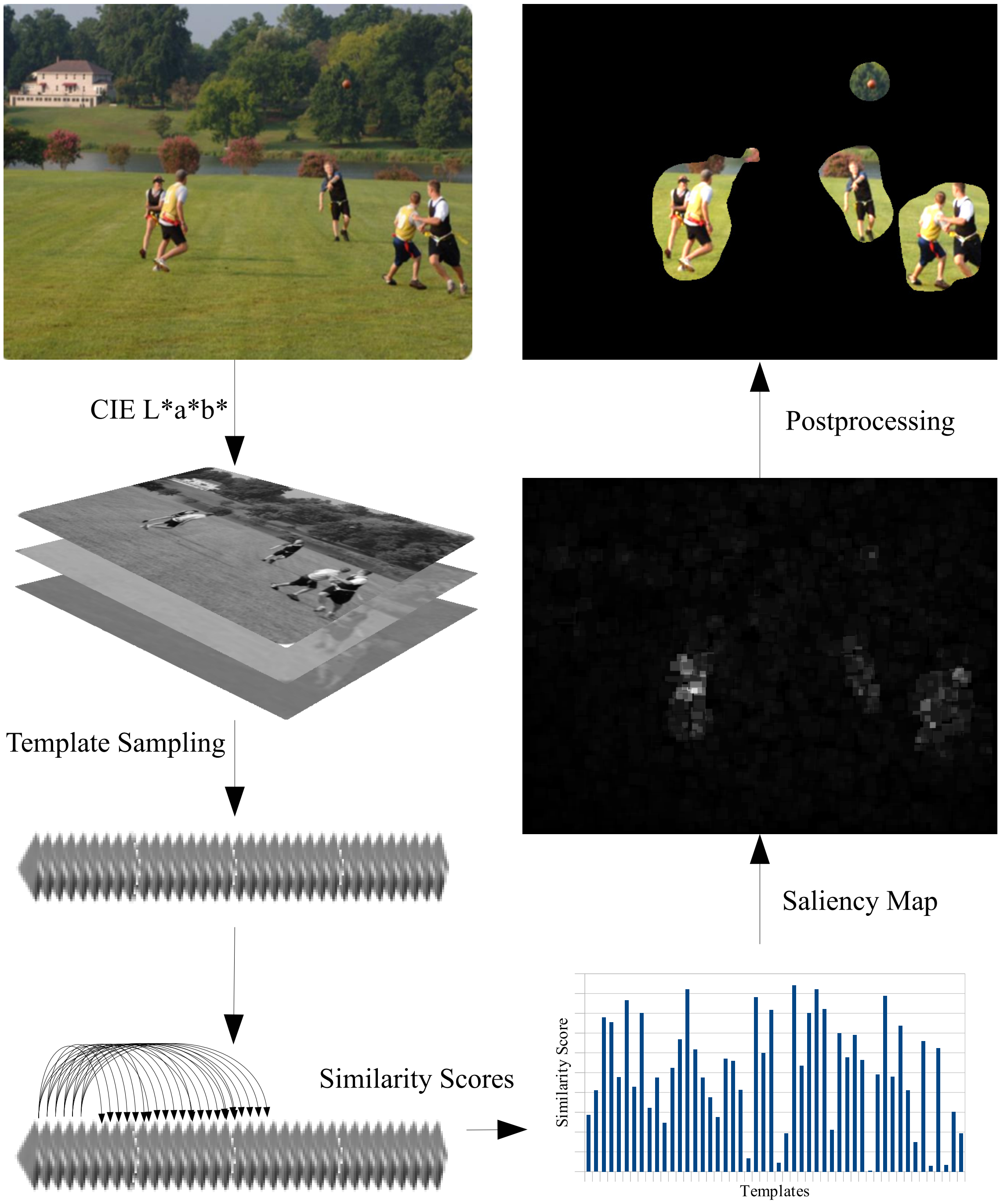


Figure 3.1 Model Overview. The input image gets converted to Lab Color Space. Then templates are randomly sampled and compared to each other using a metric which uses the L2 norm with color, shape and entropy information. Each template thereby obtains a dissimilarity score. This dissimilarity score get back-projected to the position in the image where the template was sampled from. The higher the score the more unique the template and the more salient this region is. Using simple smoothing, morphological operators and thresholding, the saliency map can be further post-processed to obtain better results.

Algorithm 1: Sampled Template Collation for Saliency Maps

```
Data: Image I; Set of templates T; Sampling rate n; Saliency Map S
Result: Saliency Map S
for i ← 1 to n do
    getRandomPosition (p_i);
    t_i = sampleTemplateFromImage(p_i, I);
    addTemplateToSet(t_i, T);
end
forall the t_i ∈ T do
    forall the t_k ∈ T ≠ t_i do
        s = calculateSimilarityScore(t_i, t_k);
    end
    setSimilarityScore(s, t_i, S, p_i);
end
```

calculated faster and are useful for generating single fixation points, more templates give a finer resolution and a more accurate and complete saliency map.

3.2 Collation Calculation

After the sampling process, each template T is compared with each other template of the same size. Different characteristics can be used to calculate the differences between the templates. For our evaluation we used the features *color space, shape, distance and entropy*. The model can easily be extended to take different and more complex measures into account, like for example the correlation coefficient or a higher weight for templates which might contain faces using simple template matching.

3.2.1 Color Space and Shape

Different color spaces like RGB or HSV were tested, CIE Lab provided us with the most consistent results. CIE Lab is a three-dimensional color-opponent space with L representing the lightness and a and b the color-opponent dimensions. A benefit is that the CIE Lab color space is perceptually uniform, which means that a change in color values should produce a change of the same visual importance. We use a L_2 norm to calculate the pixel-wise differences of lightness L and color-opponent dimensions a and b between two templates T_1 and T_2. The pixel-wise operation incorporates differences in shape into the dissimilarity score measure.

$$
\begin{aligned}
l &= ||T_{1_L} - T_{2_L}||_{L_2} = \sqrt{\sum (T_{1_L} - T_{2_L})^2} \\
a &= ||T_{1_a} - T_{2_a}||_{L_2} = \sqrt{\sum (T_{1_a} - T_{2_a})^2} \\
b &= ||T_{1_b} - T_{2_b}||_{L_2} = \sqrt{\sum (T_{1_b} - T_{2_b})^2}
\end{aligned}
\tag{3.1}
$$

3.2.2 Distance Weight

We include a distance weight to the dissimilarity score to account for local salient areas. Templates which are closer together have a higher weight than templates which are e.g. on the opposite side of the image. We compute the distance weight w by

$$w = 1 - \frac{d(T_1, T_2)}{max(d)} \tag{3.2}$$

with $d(T_1, T_2)$ being the euclidean distance between the pixels' positions in template T_1 and T_2. The maximum possible distance $max(d)$ is the diagonal of the image. We set the distance weight to zero, if $d(T_1, T_2)$ is above a certain threshold (in our case half the maximum distance), this immensely improves computational performance while having no impact on the overall AUC(ROC) score.

3.2.3 Entropy

There exist numerous visual attention models which are built on information theoretic foundation to find the most salient areas [Bruce and Tsotsos, 2005; Lin et al., 2010; Tamayo and Traver, 2008].

Entropy is a measure for information. A low entropy value means there is only little information carried in the template, high entropy means high information. We calculate the entropy in regard to the pixels intensity distribution in a template. A low result means in our case a template with almost the same pixel intensity at all positions, whereas a high entropy would be a uniform distribution of intensities. This entropy approach introduces a simple position and rotation invariant descriptor for texture. Figure 3.2 visualizes a template with low entropy at the bottom and a template with a high entropy at the top. A more uniform distributed color histogram, like in the upper case means a higher entropy. A more narrow distribution like in the bottom case means a lower entropy.

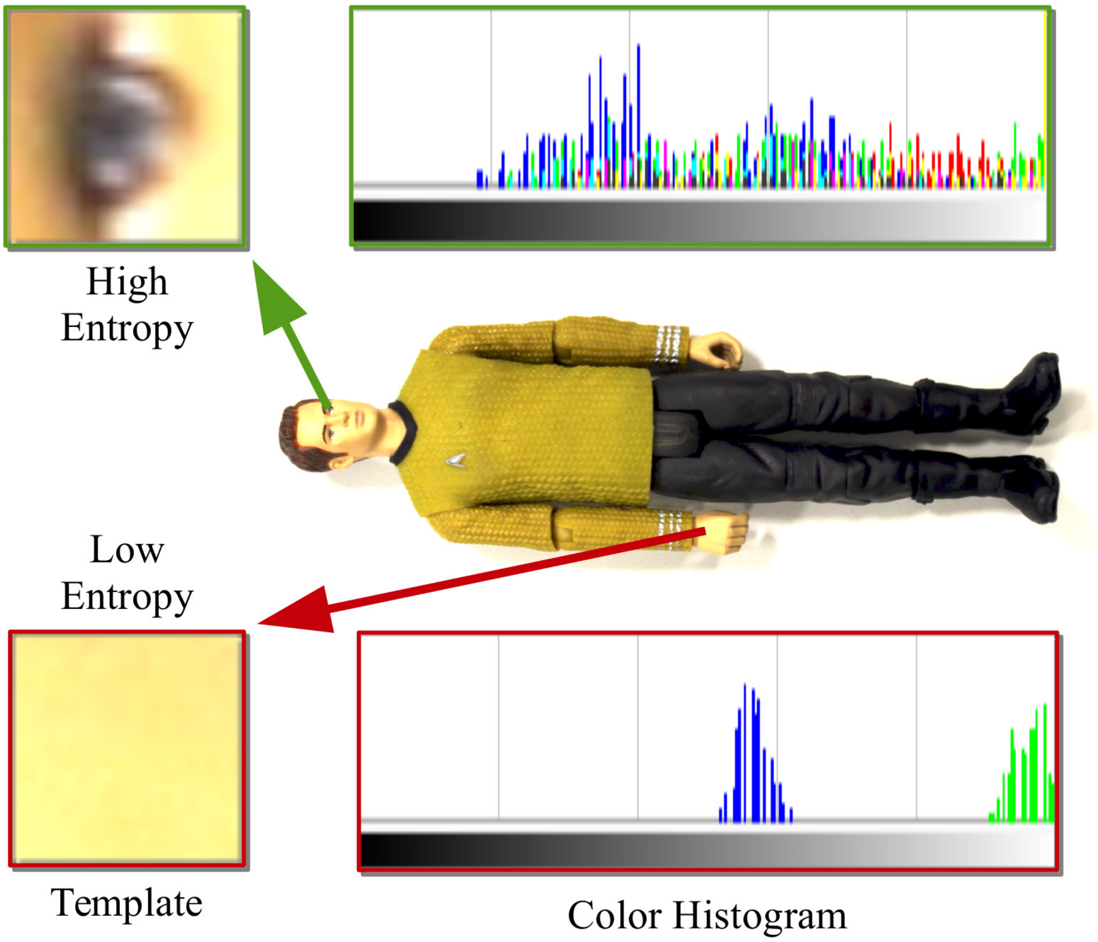


Figure 3.2 Two templates sampled from different locations. The upper template contains the eye and has a high entropy because of a broadly distributed color histogram. The bottom template was sampled from the hand and has a very narrow color distribution, which results in a low entropy.

We integrate the self-information of a template in our model by using:

$$H(X) = -\sum_{m=1}^{M} p_m \log p_m \tag{3.3}$$

with p_m being the relative frequency of brightness value m within the template. Using entropy we gain slightly better results (see table 3.2), as areas which would be salient because of their lightness and color uniqueness - e.g. a small area of a blue sky in the top of an image are not necessarily salient to a human subject.

We finally calculate the overall dissimilarity score s by calculating:

$$s = l(a + b) * w * H(T_1)H(T_2) \tag{3.4}$$

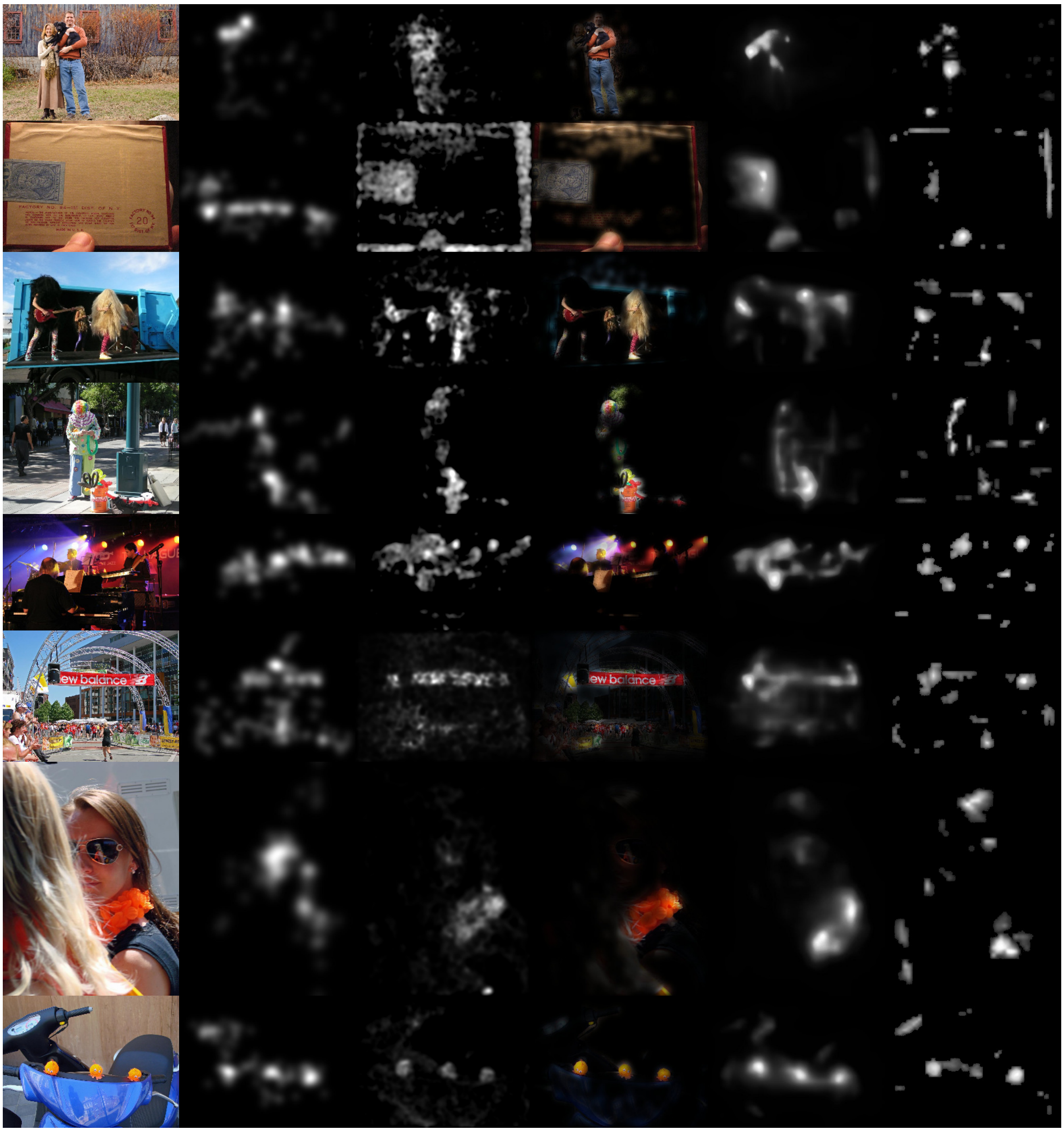
3.3 Evaluation

The model was evaluated using two open accessible saliency benchmark databases. These databases measure the similarity between the model and human observers. We also evaluated our model for computational efficiency during online processing and the effect of the sampling rate on the map stability.

3.3.1 Saliency Benchmarks

We tested our approach on Judd's et al. [Judd et al., 2012] saliency benchmark database[1]. The database contains 300 natural images with eye tracking data from 39 observers. Including a center bias our model performs significantly better than without a center bias (see table 3.1). The best results were achieved calculating the saliency map with $0.6 * centermap + 0.4 * our\ model$. The center map is a symmetric Gaussian stretched to fit the aspect ratio of the image. The factors were optimized using a different training set - see [Judd et al., 2012] for more details on center map and the optimization. Without a center bias our model still outperforms standard models like Itti & Koch (see table 3.1). See figure 3.3 for a comparison of images and saliency maps for several models.

[1] http://people.csail.mit.edu/tjudd/SaliencyBenchmark/



(a) Original Image (b) Human Fixation Map (c) **Our Model** (d) Intersection /w Input (e) GBVS (f) Itti

Figure 3.3 Sample images and saliency maps for several models. Column (a) shows the input image, column (b) the fixation map of multiple human subjects. Column (c) shows the generated saliency map using our Sampled Template Collation model. Column (e) shows the results using Graph-based Visual Saliency [Harel et al., 2006] and column (f) the ones from Itti's model [Itti and Koch, 2001] (Images taken from the MIT saliency benchmark database [Judd et al., 2012].)

Table 3.1 Results of Judd's et al. saliency benchmark dataset. Our model (blue) outperforms Itti & Koch's model even without center bias (/wo CB) and performs similar to GBVS with center bias (/w CB)

Model	**ROC**	**Similarity**
Deep Gaze 2 [Kümmerer et al., 2014]	0.870	0.460
GBVS [Harel et al., 2006]	0.801	0.472
Sampled Template Collation /w CB	0.794	0.477
Multi-Resolution AIM [Advani et al., 2013]	0.772	0.471
Center Based	0.783	0.451
Sampled Template Collation /wo CB	0.687	0.357
Torralba [Torralba et al., 2006]	0.684	0.343
Itti & Koch [Itti and Koch, 2001]	0.562	0.284
Chance	0.503	0.327

We also tested our model with the ImgSal database[1] [Li et al., 2013], which contains 235 color images, divided into six different categories ordered by their salient region size. We achieve an overall AUC(ROC) score of about 82% using a center bias, see table 3.2.

Our model outperforms state-of-the-art models like Multi-Resolution AIM [Advani et al., 2013] or long standing Itti et. al's [Itti and Koch, 2001]. There are models which outperform our system, like Judd et al. [Judd et al., 2012], which train a model using human fixation data, which incorporates the human's strong attention focus towards faces, persons or animals. We are not explicitly aiming at an adaption of the human fixation, but rather for generating a fixation point or region for salient areas. Our model however can be easily extended to bias templates with e.g. faces over templates with no faces using simple template matching.

Figure 3.4 and figure 3.5 present more challenging tasks for visual attention systems. The first image shows a landscape with a small vehicle in the right middle. Our model is able to detect the vehicle as most salient point in the scene. The second image shows a person camouflaged with the background. Although the color and texture is identical to the rest of the image, the person clearly is detected as the most salient area.

[1] http://www.cim.mcgill.ca/ lijian/database.htm

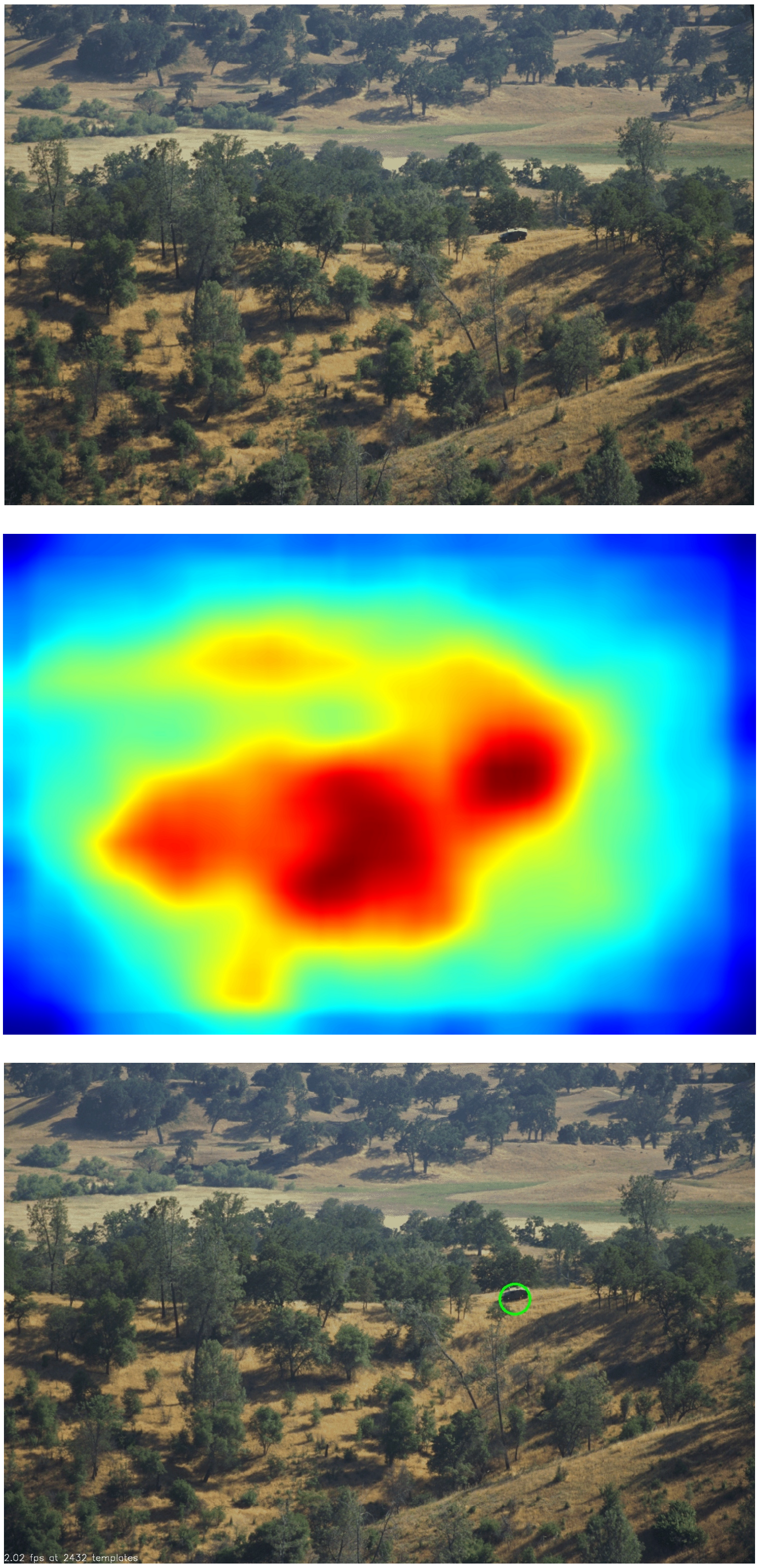


Figure 3.4 Sample image from a database of vehicles in natural background [Itti, 2000]. The image was processed using the sampled template collation method to create the saliency heat map. The generated most salient point is located on the vehicle (circled in green).

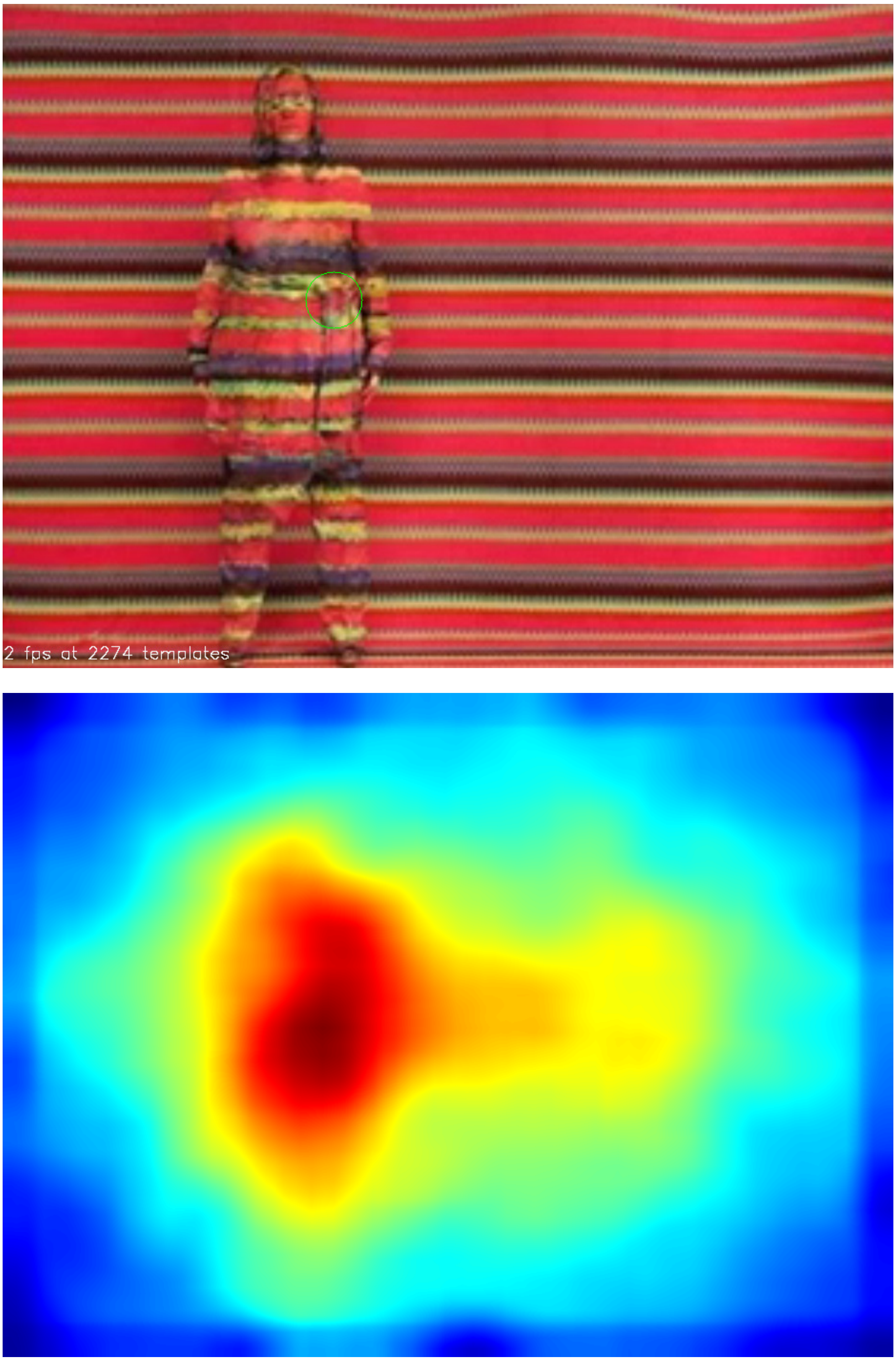


Figure 3.5 Test image with uniformly distributed texture. The saliency map was generated using our sampled templates collation approach. The irregularities in the image are detected as most salient area, although the person is camouflaged with the texture identical to the background.

Table 3.2 Results of ImgSal saliency benchmark dataset with and without template entropy (TE); without center bias (CB) and without smoothing (Sm.).

Category	/w TE	/wo TE	/wo CB	/wo Sm.
Large	0.819	0.818	0.793	0.707
Intermediate	0.813	0.799	0.780	0.670
Small	0.817	0.790	0.772	0.669
Cluttered	0.808	0.808	0.780	0.677
Repeating Distr.	0.848	0.844	0.816	0.715
Large & Small	0.826	0.809	0.798	0.706
Overall	**0.818**	**0.805**	**0.785**	**0.690**

3.3.2 Frame Rate Control

In the context of real-time processing it is important to be able to adaptively react to different computational scenarios and to maintain a certain degree of low-latency computation. Our model's main aspect is the sampling process which has the major benefit, that it can be adjusted online. To estimate the computational speed of our performance we adaptively change the number of sampled templates to match a standard camera image frequency of about 30 fps at 640x480 pixels. If the processing is slower than 30 fps, less templates are sampled; if faster, more are sampled. We tested this setting on an intel i7 with 3.4 GHz and were able to sample about 130 templates using one core and about 440 templates using four cores for every camera frame captured at 30 Hz.

We enhanced our approach to dynamically adapt the sampling rate to achieve a desired frame rate using the following equation:

$$samples_{new} = \sqrt{\frac{fps_{current}}{fps_{desired}}} * samples_{current} \tag{3.5}$$

which assumes a complexity of $O(n^2)$ for the worst case scenario with no distance weight. By reducing the sampling rate, the frame rate can be kept constant even if computationally intensive programs run on the same computer. Figure 3.6 shows how the frame rate control works during a test run. The high framerate in the beginning is reduced by increasing the sampling rate using equation 3.5 to match a desired frame rate of 15Hz. The frame rate control can easily be extended to incorporate more complex constraints like a minimum desired sampling rate.

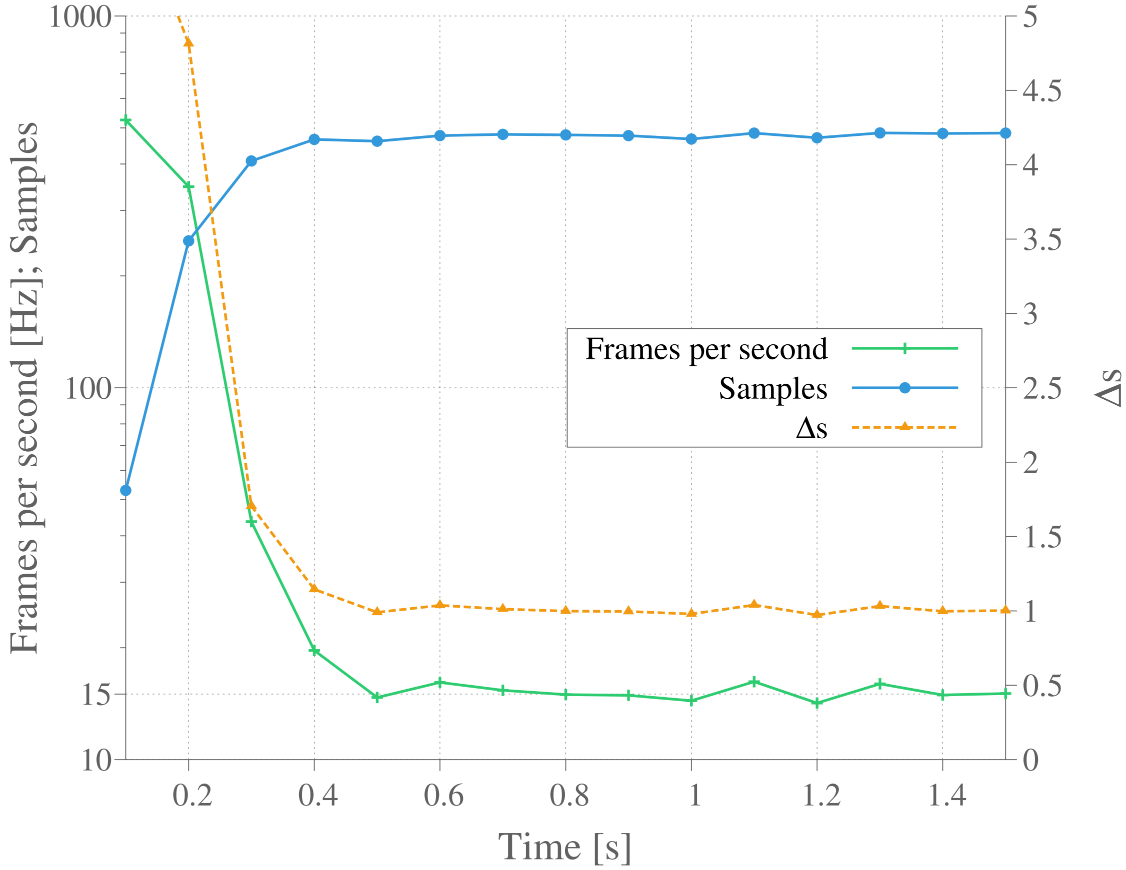


Figure 3.6 Number of samples and framerate during a test run. The plot shows the results of a test run of the visual attention system with a previously defined desired framerate of 15 Hz. In the beginning the framerate is higher than anticipated, so the sampling rate is increased until the desired framerate is achieved. The Δs represents the adjusted sampling factor, which is constantly calculated depending on the current framerate and sampling rate.

3.3.3 Sampling

We evaluated the effects of the sampling process on the consistency of the saliency map and the salient point position. Due to the random selection of the template position during the computation, the generated saliency maps of two successive frames are different from each other. Depending on the number of sampled templates this effect has varying degrees of impact on the generated saliency maps.

The more templates are sampled, the less the deviation between the maps and the higher the stability of a generated saliency map. We measure the deviation by calculating the L1-norm of two generated saliency maps of the same input image. The deviation of the most salient point is measured by the euclidean distance between the two points in the image. The results are displayed in figure 3.7. Both values are normalized, so that 100% deviation means the maximal possible deviation. From about 100 sampled templates, the deviation in the saliency map and the most salient points are constant with about $0.2 * 10^{-3}\%$ and 4% deviation, respectively. Figure 3.8 visualizes the different saliency map generated from the same image but with a low sampling rate ($\approx$ 20 templates) from two successive frames. The low density of distributed templates is the reason for missing salient areas like the green pen in the image at the bottom.

A beneficial side effect of the deviation induced to the randomness in sampling are the different calculated position of most salient points. Instead of a static result, the maxima jump over other similar salient points, which leads to a visual scan path. In other visual attention models, this scan path was created calculating a winner-take-all network to find the next most salient point. This computational expensive step can be avoided.

Figure 3.9 shows an example of two different maximum salient points calculated from the same image with a high sampling rate of 400 templates. Although the resulting maps are similar to a certain degree, the most salient point lies alternating on the green pen (top image) or yellow image (bottom image), because both areas are similar salient.

Our approach can be extended to account for lower salient areas, by keeping templates sampled from the most salient area and assigning them a lower weight for the STC calculation in the next frame. This has the advantage, that instead of using old image data like in the winner-takes-all approach, a new acquired image can be used for generating the saliency map. This approach would have two advantages:

1. It wouldn't matter if the position of most salient object would change, as the templates sampled from the new position would get lower weighted as well.
2. It would take new objects into account which haven't been in the scene before.

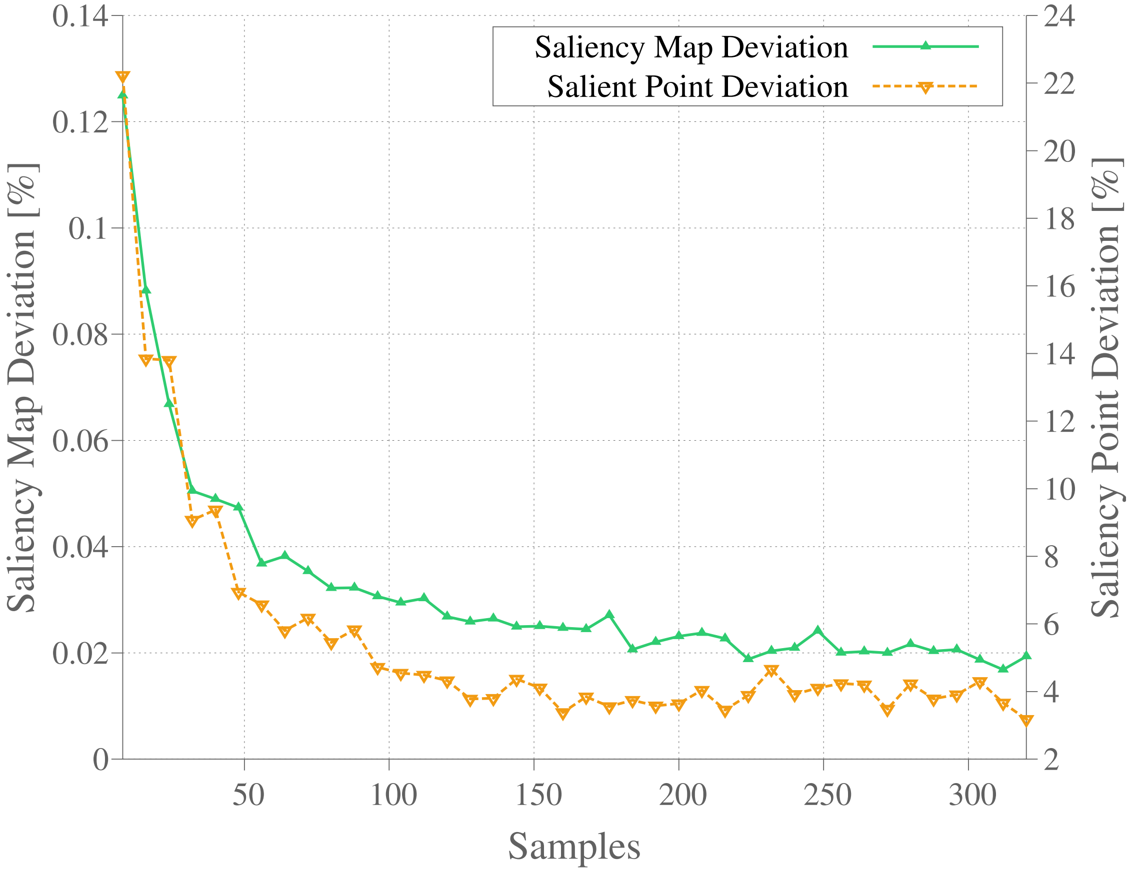
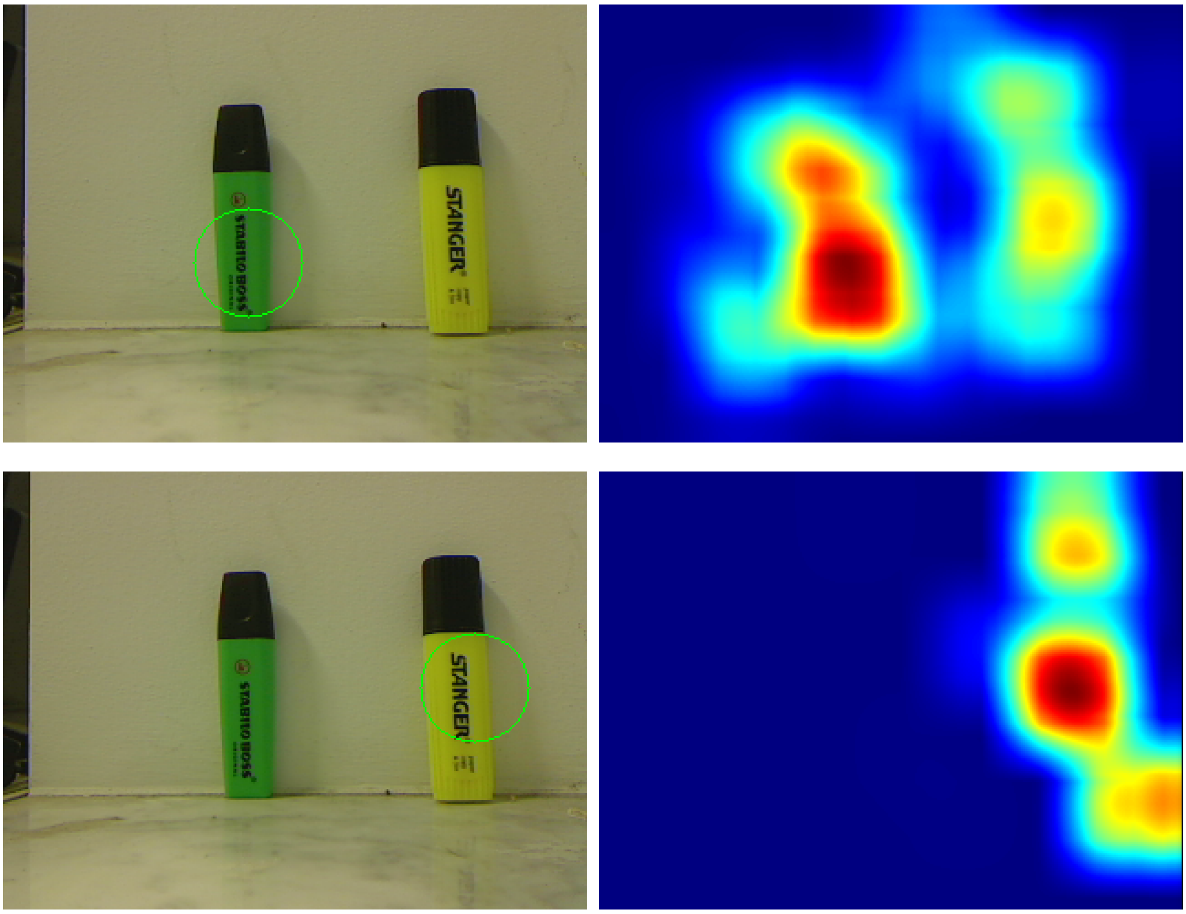


Figure 3.7 Saliency Map Stability. The graph shows the deviation in percent between saliency maps generated with increasing sampling rates (green line) and the distance deviation of the most salient point in those saliency maps (yellow line). While in the beginning the resulting saliency maps clearly differ from each other, the deviation remains constant at a sampling rate of about 100 templates. The map deviation percentage was calculated using the maximum possible deviation, which is size of the image times maximum saturation value. In case of the salient point deviation percentage, the maximum possible distance was used which is the image diagonal.

3.3.4 Computational Performance

Our model's main aspect is the sampling process which has the major benefit, that it can be adjusted online. To describe the complexity, we use the Big O notation, which characterizes functions by giving an upper bound according to their growth rates depending on the input values. After the sampling process, each template is compared to the other templates by calculating a dissimilarity score. Usually this results in a number of function calls of n templates times n templates, so a complexity of $O(n^2)$. In our case

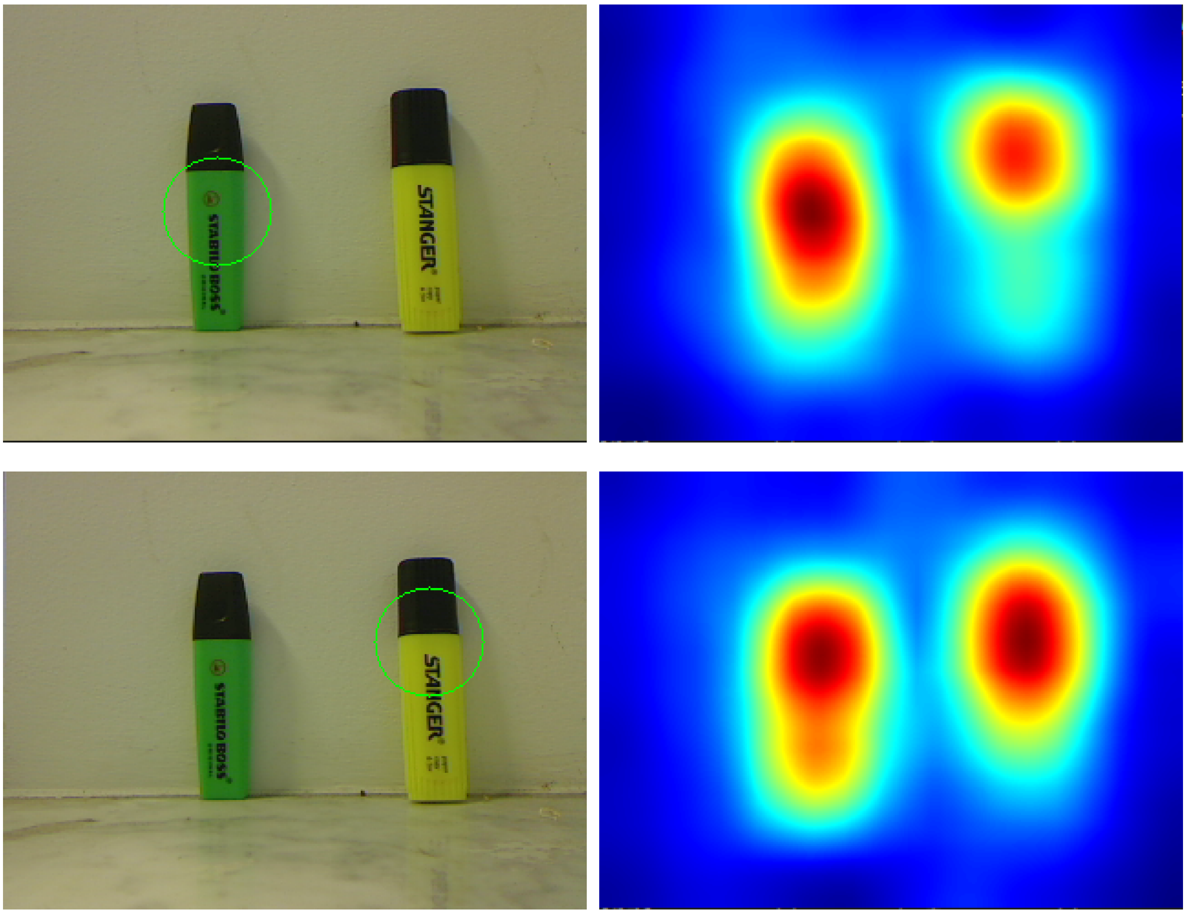


(a) Image with most salient point (b) Saliency Map

Figure 3.8 Difference in saliency maps with a low sampling rate. The saliency maps were generated with a very low sampling rate (≈ 20 templates) from two successive frames which show the same image. The produced results vary significantly, due to the low density of templates, which makes it more likely to miss salient areas, like the green pen in the bottom right image. This behavior can be measured by the saliency map deviation (see figure 3.7).

the used dissimilarity score is a commutative function, so that $f(T_1, T_2) = f(T_2, T_1)$. So for the first template it takes n function calls, for the second template $n-1$, for the third template $n-2$ and so on. This leads to a complexity of $O(\frac{1}{2}n(n+1))$, which for a sampling rate of e.g. 100 templates would require 5.500 function calls, compared to 10.000 for non-commutative functions.

The complexity can be further reduced by introducing a distance threshold, which only allows close templates to be compared (see 3.2.2). We approximate the system's complexity by assuming that we calculate the k nearest neighbors of each template, which



(a) Image with most salient point (b) Saliency Map

Figure 3.9 The saliency maps were generated with a higher sampling rate ($\approx$ 400 templates) from two successive frames which show the same image. The produced results are similar but have different maximum salient points. This fluctuation has the advantage of avoiding further processing for example a winner-take-all computation to find other similar salient points. Other salient points can automatically be detected to a certain degree just by exploiting the effects of random sampling.

has the complexity $O(n \log n)$. The number of dissimilarity score calculations therefore is $n * k$. The complexities combined are

$$
\begin{aligned}
& O(n \log n) + O(n * k) \\
& = O(n \log n) + O(n) \\
& = O(max(n \log n, n))
\end{aligned}
\tag{3.6}
$$

The inequation $n \log n > n$ is true for all $n > 2$. As the number of sampled templates will always be larger than 2 for an operating system, we can approximate that the overall complexity is $O(n \log n)$:

$$\forall n > 2 : O(max(n \log n, n)) = O(n \log n). \tag{3.7}$$

We evaluated the computational complexity on the basis of achieved framerate per sampling rate. Therefore we steadily increased the number of samples while calculating the resulting framerate (see figure 3.10). The generated curve (green line) shows an approximation to our expected complexity of $O(n \log n)$ (blue line). The CPU usage (orange line) increases linear with the number of samples. As shown in figure 3.7, the generated saliency maps remain stable at about 100 samples, which would give us a framerate of about 140 Hz with a CPU Usage of 60%.

We perform about as good as GBVS [Harel et al., 2006] in regard to the ROC Score for predicting the human gaze. Our approach however is scalable and we have a much lower complexity of $O(n \log n)$ compared to $O(n^4 K)$.

In case of a non-commutative function we propose algorithm 2: Neighbor-based Sampled Template Collation for Non-Commutative Dissimilarity Score Functions. Instead of comparing a set of templates, which would have a complexity of $O(n^2)$, we sample a template and then sample k neighbors around that template within a radius r and calculate the dissimilarity score for that template. This similarity score is then added to the saliency map. The k neighbors only are used with one template and then discarded. We repeat that step until we have sampled all n templates and calculated the dissimilarity scores.

This approach has the complexity $O(n \times k)$. For a sufficiently small k (namely $k < n$), this approach has a lower complexity than the original approach.

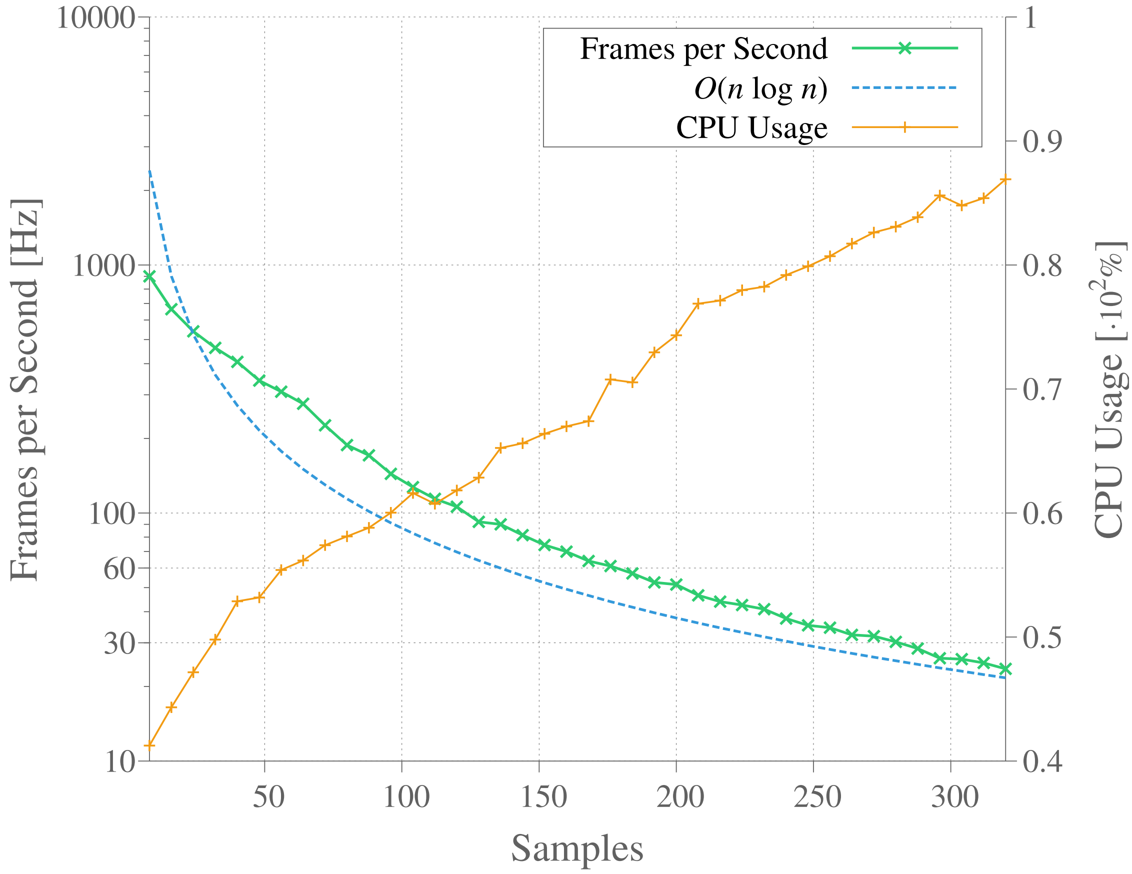


Figure 3.10 Processing performance measured in frames per second when constantly increasing the sampling rate. We mathematically approximated the complexity of our system to be $O(n \log n)$ and empirically approximate the real complexity with the measured framerate against the sampling rate. The corresponding graphs show that the approximations are similar. As previously shown, the saliency map deviation remains stable at about 100 samples which would give us a framerate of about 140 Hz with a CPU usage of 60%. Therefore we can adapt the sampling rate to match the camera's framerate.

Algorithm 2: Neighbor-based Sampled Template Collation for Non-Commutative Dissimilarity Score Functions

```
Data: Image I; Set of templates T; Sampling rate n; Number of Neighbors: k;
      Radius r; Saliency Map S
Result: Saliency Map S

for i ← 1 to n do
    getRandomPosition (p_i);
    t_i = sampleTemplateFromImage(p_i, I);
    for m ← 1 to k do
        getRandomPositionWithinRadius (p2, r);
        t2 = sampleTemplateFromImage(p2, I);
        s = calculateSimilarityScore(t_i, t2);
        addSimilarityScoreToMap(s, p_i, S);
    end
end
```

3.4 Summary

In this chapter we presented Sampled Template Collation - a method for visual attention and saliency. The benefits of STC are the scalability and computationally efficiency. Our method outperformed state-of-the-art in terms of prediction accuracy of human observers and in terms of computational complexity. Sampled Template Collation is especially suited for the applications in time-crucial scenarios and test cases with limited hardware, like a mobile platform or a humanoid robot.

In the next chapter we present the application of Sampled Template Collation for Object-based attention. A biologically-inspired method based on visual attention for segmenting objects.

Chapter 4

OBJECT-BASED ATTENTION USING SAMPLED TEMPLATE COLLATION

Object-based attention explains the behavior of neural responses when an object is fixated. Visual stimuli are adjusted in favor of the particular object, which enhances the processing of the object's features. So far, only little research has been conducted towards the application of object-based attention in technical applications.

In this chapter, our object-based attention system is presented. The method we developed is based on our sampled template collation model. The advantages are the low computational complexity and the improved object recognition results due to the segmentation. The first section describes how sampled template collation is used for object segregation. The second section presents the visual segmentation results and the improved object classification performance.

Object-based attention describes the relationship between a fixated object and the visual stimuli in the visual cortex. Desimone and Duncan describe two basic phenomena that define the problem of visual attention [Desimone and Duncan, 1995]. The first one is the limited capacity for processing the information available on the retina. The second one is the ability to filter out currently unnecessary information, which enhances the visual representation of objects, even if spatially occluded in cluttered real-world scenarios. They suggest that the quality of sensory representation of a fixated object is improved. This results in an enhanced processing of the object's features.

Object-based attention suggests a pattern-specific attentional filtering in the visual cortex. Activity patterns in early visual areas are strongly biased in favor of the attended object [Cohen and Tong, 2013]. This phenomenon contributes towards the recognition of objects in higher cortical areas [Walther et al., 2005; Walther and Koch, 2006].

4.1 Sampled Template Collation for Object-Based Attention

Our object-based attention approach is based on Sampled Template Collation and preserves the same advantages like scalability and efficiency.

Figure 4.1 shows our processing pipeline. First we resize the image - we experienced the best results with an image size of $160x120$. We apply Gaussian blurring to reduce noise and texture, then dilate and erode the image, which helps with the isolation of individual elements and joining smaller separated elements on an object. We convert the image from RGB to Lab color space and calculate the object-based attention map using Sampled Template Collation.

One single seed template is taken from the area with the highest salient point computed by our visual attention procedure. All following sampled templates are then compared to this seed template using a similar metric as in equation 3.4:

The color dimensions a,b and lightness l are calculated using a L_2-norm:

$$
\begin{aligned}
l &= ||T_{1_L} - T_{2_L}||_{L_2} = \sqrt{\sum (T_{1_L} - T_{2_L})^2} \\
a &= ||T_{1_a} - T_{2_a}||_{L_2} = \sqrt{\sum (T_{1_a} - T_{2_a})^2} \\
b &= ||T_{1_b} - T_{2_b}||_{L_2} = \sqrt{\sum (T_{1_b} - T_{2_b})^2}
\end{aligned} \tag{4.1}
$$

The entropy H is calculated with

$$H(X) = -\sum_{m=1}^{M} p_m \log p_m \tag{4.2}$$

with p_m being the relative frequency of brightness value m within the template.

The final dissimilarity score s is

$$s = (a + b) + (\alpha * l) + |H(T_1) - H(T_2)| \tag{4.3}$$

with $\alpha = \frac{1}{3}$. Areas with a lower response are therefore more likely to contain the object. We only use a single template as seed, because a set of templates with a larger spatial distribution might not contain the attended object and could lead to false object-based attention maps, especially for small objects.

After thresholding the heatmap we apply contour finding and remove all contours which don't contain the most salient point where the seed template was sampled from. This way we avoid areas which have a similar response to the center object, like the areas around the cup in figure 4.1 at "Contour". Algorithm 3 depicts the single steps involved in the generation. The method *createObjectBasedAttentionMap* describes two different approaches to the sampling process, which are shown in algorithm 4 and algorithm 5 in section 4.1.1.

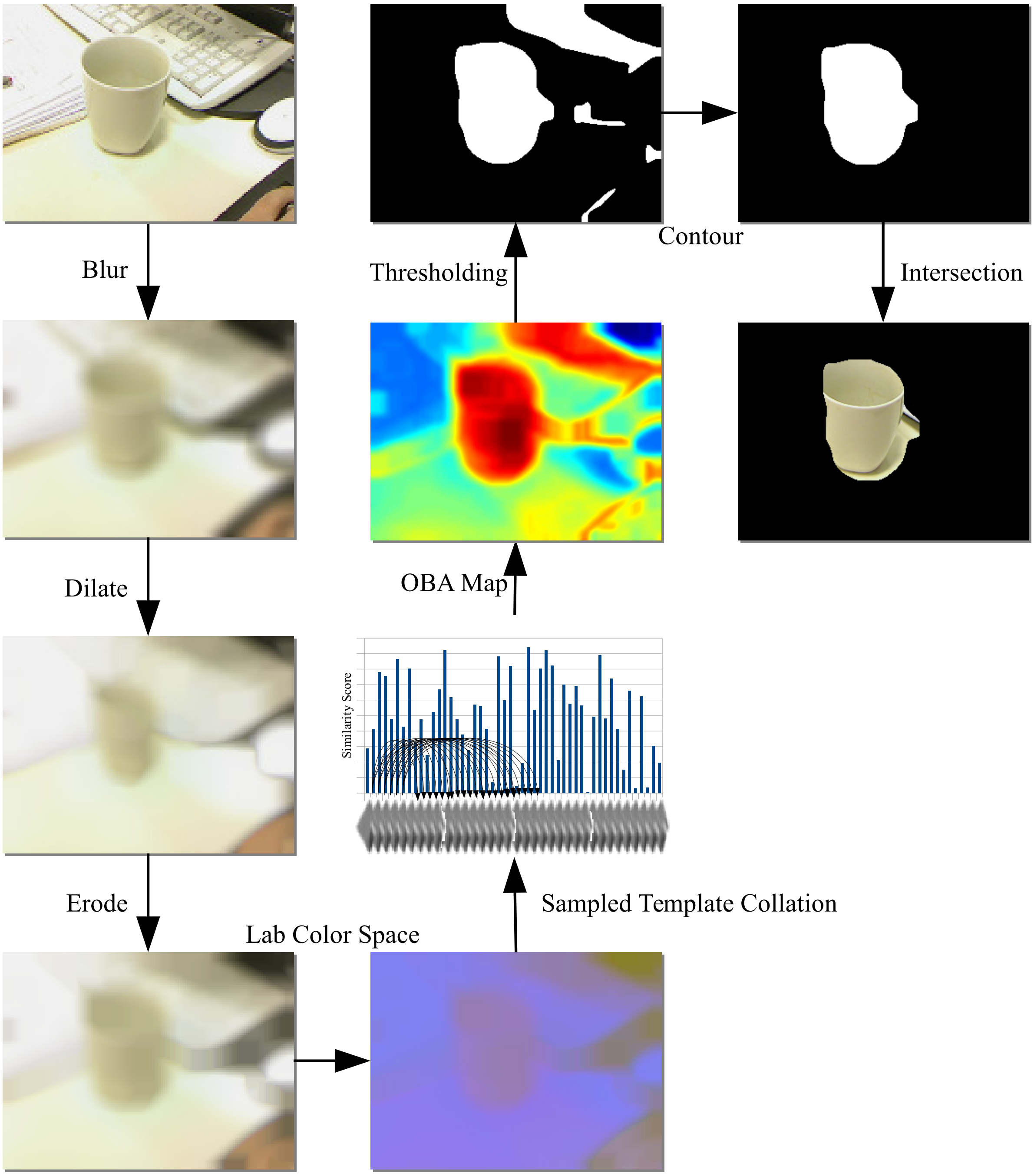


Figure 4.1 Our object-based attention model. We prepocess the image with blurring, dilation, erosion and converting it to Lab color space. Then we calculate the object-based attention map using sampled template collation. We binarize the map using thresholding and apply contour finding. Finally we remove the contours that don't contain the seed template.

Algorithm 3: Sampled Template Collation for Object-Based Attention

Data: Image I; Set of templates T; Salient Point S; Object-Based Attention Map M
Result: Object-Based Attention Map M

gaussianBlur(I);
dilate(I);
erode(I);
convertToLabColorSpace(I);
M = createObjectBasedAttentionMap(I, S) ; `/* See algorithm 4 and 5 */`
binaryThreshold(M);
findCentralContour(M);
removeOtherContours(M);

4.1.1 Dense and Sparse Sampled Template Collation

In contrast to Sampled Template Collation for visual attention, different sampling rates might have a stronger effect on the quality of the object-based attention map, as it is not guaranteed that the whole area on the object is covered during the calculation. Regions on the object might be missing in the final contour finding step, or areas that don't lie on the object might be assigned to it. Therefore we developed two different algorithms for calculating the Sampled Template Collation:

1. **Sparse Sampled Template Collation** randomly samples templates over the input image as done in the visual attention approach. We added a morphological closing operation after the heatmap is generated to cope with gaps in the representation (see algorithm 5). Figure 4.2 visualizes the different results. The images on the left side were generated without post-processing the heat map. The red dots are areas where no template has been sampled from. Those areas are later falsely recognized as contours of the object. This behavior is more likely at a low sampling rate because it can't be guaranteed that pixels in the input image are covered. These dots can be removed by applying the morphological operations dilation and erosion (see right images).

2. **Dense Sampled Template Collation** calculates the similarity at every possible position in the image. At every pixel a template is sampled and compared to the seed template (see algorithm 4). This creates a complete representation of every

pixel but is computationally more expensive than the sparse approach and not scalable during online processing.

Algorithm 4: Dense Sampled Template Collation

Data: Image I; Pixel p; Salient Point Position S; Seed Template C; Object-Based Attention Map M
Result: Object-Based Attention Map M

```
C = sampleSeedTemplateFromImage(S, I);
forall the p_i ∈ I do
    t_i = sampleTemplateFromImage(p_i, I);
end
forall the t_i ∈ T do
    s = calculateSimilarityScore(t_i, C);
    setSimilarityScore(s, t_i, S, p_i);
end
```

Algorithm 5: Sparse Sampled Template Collation

Data: Image I; Salient Point Position S; Seed Template C; Sampling Rate n; Set of Templates T; Object-Based Attention Map M
Result: Object-Based Attention Map M

```
C = sampleSeedTemplateFromImage(S, I);
for i ← 1 to n do
    getRandomPosition(p_i);
    t_i = sampleTemplateFromImage(p_i, I);
    addTemplateToSet(t_i, T);
end
forall the t_i ∈ T do
    s = calculateSimilarityScore(t_i, C);
    setSimilarityScore(t_i, M, p_i);
end
```

We evaluated the effects of the sampling by comparing the deviation between the two frames and the deviation between the two methods dense and sparse sampling. Figure 4.3 depicts the results.

The deviation between the frames is calculated by the absolute difference of the heatmaps between two successive frames, which were generated using the same input image. With a low sampling rate the deviation is low (see orange curve in figure 4.3), because the

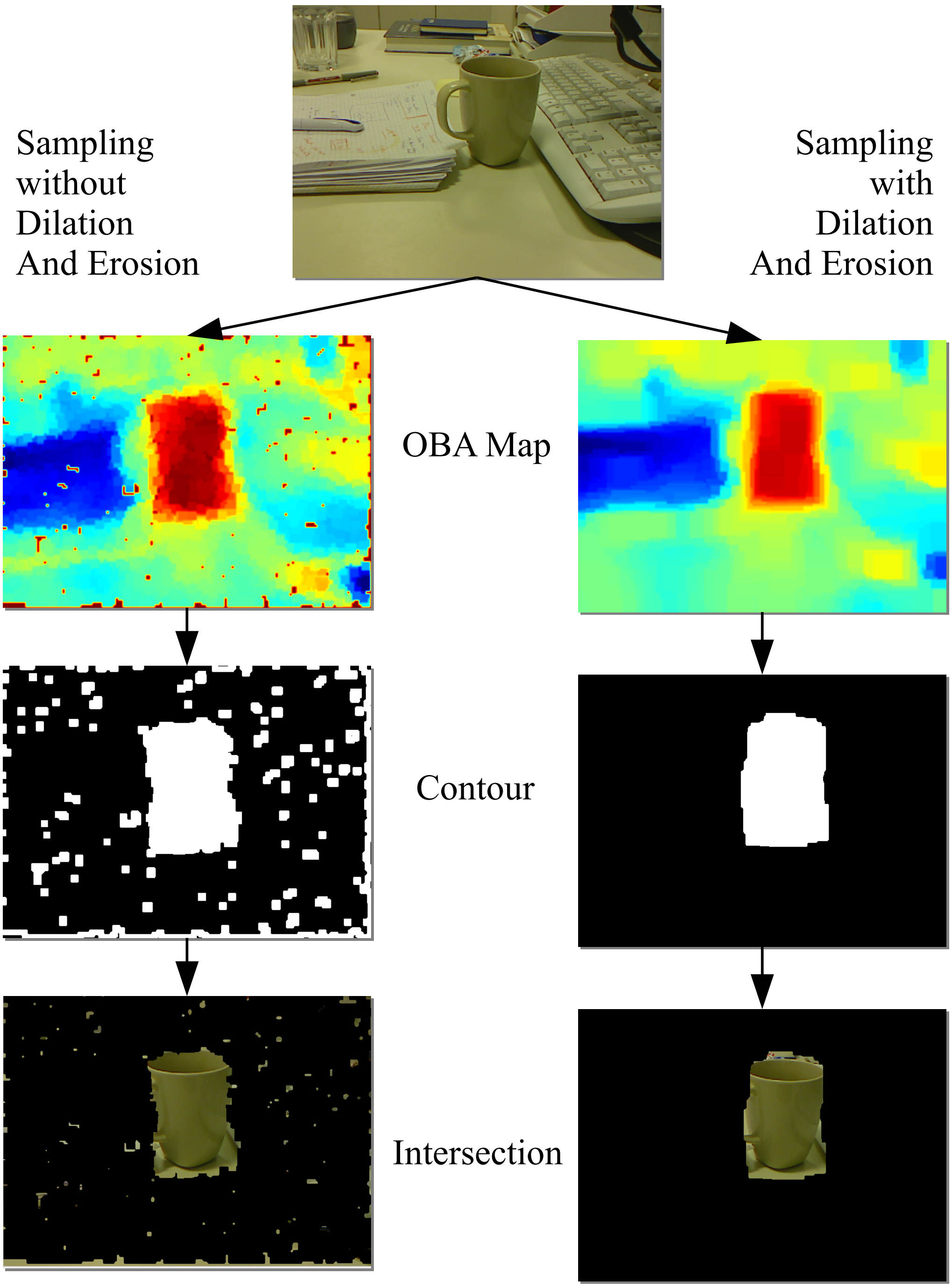


Figure 4.2 Postprocessing the object-based attention map. When the object-based attention map is generated using the sparse sampling approach, small outlier patches can occur due to the incomplete representation. This especially can happen at a low sampling rate because it is not guaranteed that all areas are covered by templates (see left column). Using the morphological dilation and erosion can help to cover up those holes (right column). Applying Contour finding and removing those contours that don't contain the most salient point can also help to get rid of the outliers.

heatmaps are initialized with a certain value and the few sampled templates cover only a fraction of the image (see figure 4.4a). The reason for the initial increase of the deviation are the random positions the templates are sampled from. With more templates, more area is covered but not enough for a constant heatmap. At about 250 templates the deviation decreases, at this point the generated heatmaps start to converge to a stable object-based attention representation (see figure 4.3c-i).

We measure the deviation between the sparse and dense Sampled Template Collation method by comparing the heatmap generated by the dense method (see figure 4.3j) and the heatmaps generated using the sparse method at different sampling rates (images a-i). We calculate the absolute difference of the heatmaps in relation to the maximal possible deviation. The deviation maximally decreases around 250 templates and stays constant at about 1000 templates similar to the deviation between frames.

The framerate achieved with the dense method is around 15Hz at a resolution of 160x120 using one CPU core. The framerate using the sparse method highly depends on the number of sampled templates. At about 1000 templates, at which point the deviation starts to converge, we get a framerate of about 240Hz. Figure 4.3 visually compares the generated heatmaps and intersections with the input image for different sampling rates. At around 1000 templates the heatmap (image g) shows little difference to the one generated with the dense approach (image j). The differences in the intersection already show good results with a sampling rate at about 300 (image c).

The presented results suggest that using the sparse sampling with a sufficiently high sampling rate produces similar good results at a much better computational performance than the dense method.

4.1.2 Efficient Sparse Sampled Template Collation

The method of sampling templates at random positions has the disadvantage that a randomly chosen position could have been visited before. If this issue is ignored and the template is sampled, the collation calculation is unnecessarily performed again which leads to an increase of redundant information and computation especially at higher sampling rates. Suppose we have N pixels, then the probability to visit an unvisited pixel is

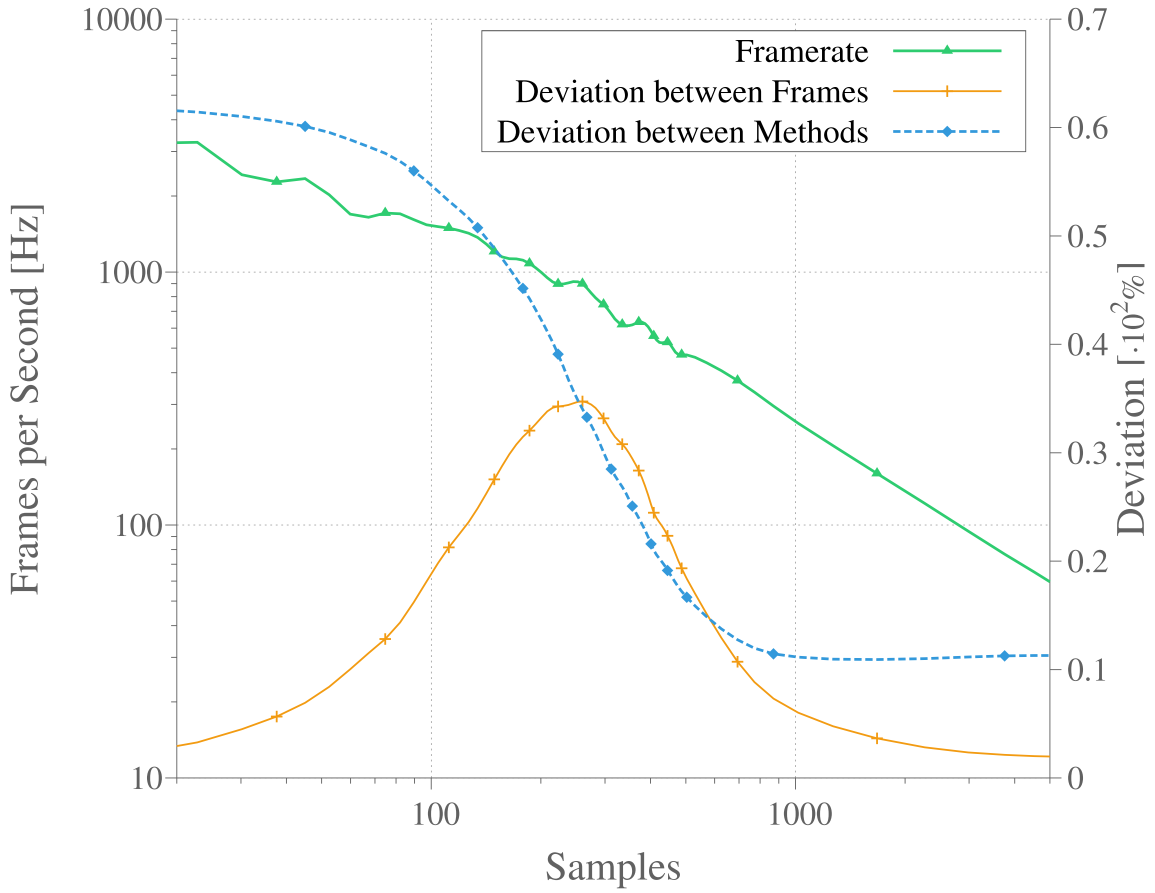


Figure 4.3 The effect of sampling on the framerate, deviation between frames and deviation between the sparse sampling method and the dense method. "Deviation between methods" displays the difference between the heatmap generated using the dense method and a currently generated heatmap using the sampling method. "Deviation between frames" displays the difference between the current generated map and the one generated a frame before that. At the beginning this deviation is low, because the heatmaps are initialized with the same value and there are only a few templates sampled which only cover a fraction of the image area. With an increasing sampling rate this deviation also increases as more area is covered but not enough for a similar heatmap. At about 250 templates the deviation decreases, at this point the generated heatmaps start to converge to a stable object-based attention representation.

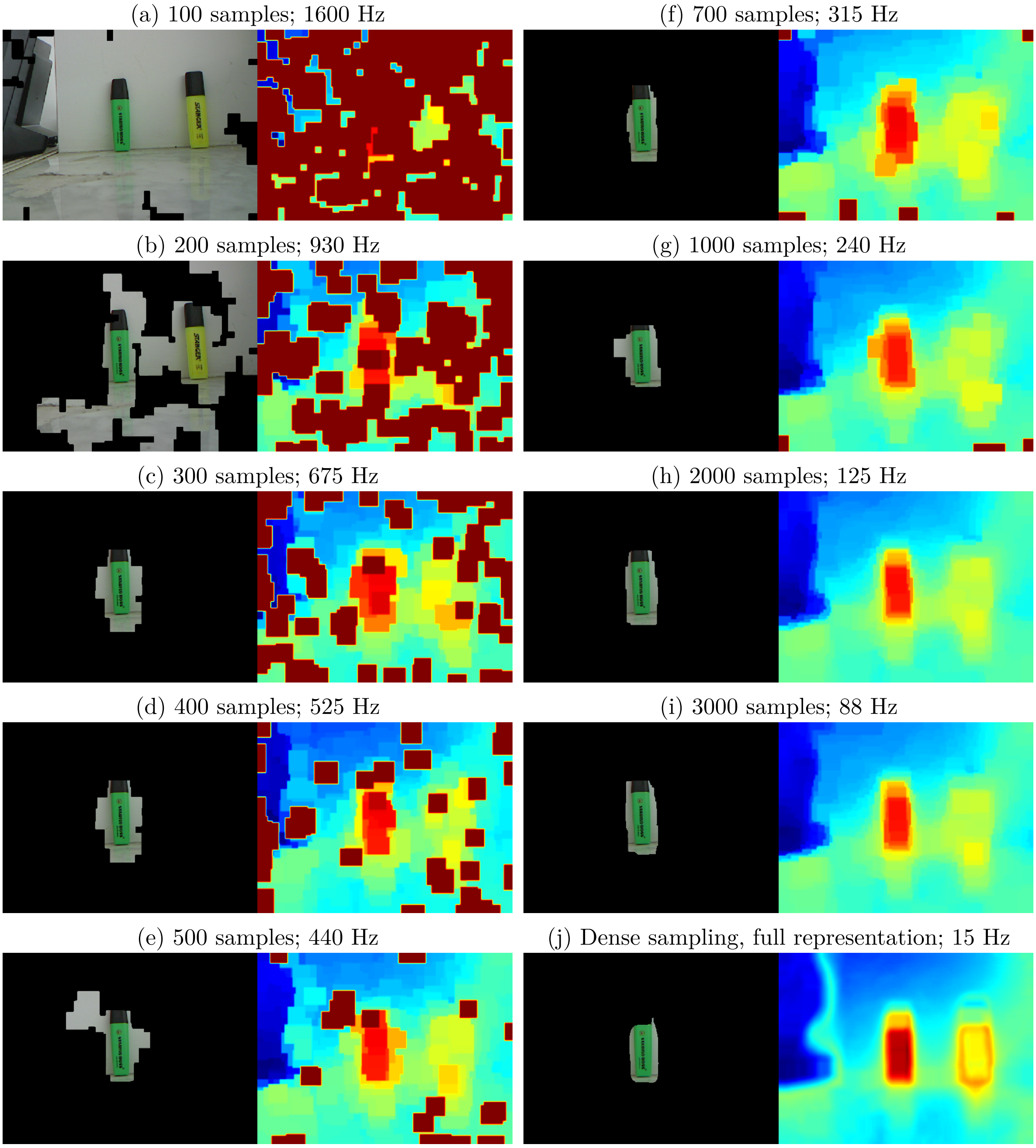


Figure 4.4 Object-based Attention Maps generated with different sampling rates. Here we visualize the effect of different sampling rates on the generation of the object-based attention heat maps and the intersection with the input image. The intersections were generated after applying thresholding and contour finding with rejecting outlier contours. We start from 100 samples in (a) to 3000 samples in (i). Figure (j) shows the map when generated with the dense method. With increasing sampling rate the maps stabilize, in the sense of the minimization of differences between frames. At the same time the deviation to the dense method also minimizes. At around 700 samples the intersection and the heat map are stable and similar to the ones generated by the dense method.

$$\frac{N}{N} = 1 \tag{4.4}$$

for the first trial. For the second trial it is

$$\frac{N-1}{N} \tag{4.5}$$

For the third trial:

$$\frac{N-2}{N} \tag{4.6}$$

and so on, until the last trial with

$$\frac{N-(N-1)}{N} = \frac{1}{N} \tag{4.7}$$

The probability for visiting an unvisited pixel after i pixels have been visited is therefore

$$p_i = \frac{N-(i-1)}{N} \tag{4.8}$$

This problem is identical to the coupon collector's problem which describes how many coupons need to be bought until the full set is complete. In our case we are interested in how many samples S it needs until all pixels have been visited. The probability p_i has a geometric distribution with expectation $1/p_i$. The overall expectation of the number of samples S necessary to visit all pixels is the sum of the single expectations of p_i:

$$\begin{aligned} E(S) &= \frac{1}{p_1} + \frac{1}{p_2} + \ldots ... + \frac{1}{p_n} \\ &= \frac{n}{n} + \frac{n}{n-1} + \ldots ... + \frac{n}{1} \\ &= n \times (\frac{1}{1} + \frac{1}{2} + \ldots ... + \frac{1}{n}) \\ &= n \times H_n \end{aligned} \tag{4.9}$$

with H_n being the harmonic number. For an image of size $40x30$ with $40x30 = 1200$ pixels the expected number of samples to be taken until all pixels have been visited is 9201. For an image with $160x120 = 19200$ pixels it already takes 200.446 samples. This clearly is computational highly inefficient if a complete representation is desired.

Algorithm 6 describes a method which avoids generating any redundant information. First a list of unvisited points is created from the input image. In our case this list contains all points in the image but could also be adjusted to for example every second point only. During the sampling a random point in this list is chosen for sampling a template from this position. The point is then removed from this list of unvisited points. This approach guarantees that every point is visited only once at most and that every additional template sample increases the information and density in the object-based attention map.

Figure 4.5 visualizes the development of the inefficient sparse template sampled collation approach of visited pixels for an image of size $40x30$ with 1200 pixels. The green curve represents the percentage of visited pixels, which has only about 60% visited pixels at 1200 samples. Figure 4.6 display the results for the efficient sparse sampled template collation approach. The green line is linear and direct proportional to the number of sampled templates. The maximum number of the 1200 possible positions is reached in 1200 sampling steps. Here the probability to visit an unvisited pixel is always 1 (see orange curve). In contrast the probability of the inefficient approach decreases with the number of samples (figure 4.5 orange curve). At about 800 samples half of the pixels have been visited, so for every visited pixel $800/600 = 1.33$ pixels had to be sampled. The probability to find an unvisited pixel is already at only 50%.

Postprocessing the map with morphological operators like explained in the previous section can help to cover unvisited pixels. The blue curve in figure 4.5 shows the substantial increase in visited pixel when applying postprocessing. At about 1000 samples almost all pixels are covered. The efficient sparse template sampled collation approach can also benefit from postprocessing see figure 4.6). The blue curve shows that the whole image is covered at about 600 samples.

Concluding we can say that the efficient sparse sampling template collation approach has shown to be superior in every aspect to the previous version. Using a list of unvisited points avoids generating redundant information, it is more computationally efficient and generates denser object-based attention maps. In addition it maintains the randomness of sampling which has previously been shown to have beneficial properties in online processing.

Algorithm 6: Efficient Sparse Sampled Template Collation

Data: Image I; Non visited Points P; Salient Point Position S; Seed Template C; Sampling Rate n; Set of Templates T; Object-Based Attention Map M;
Result: Object-Based Attention Map M

```
V = createSetOfNonVisitedPoints(I);
if n > sizeOf(V) then
    // The sampling rate n is adjusted, if it is higher than the
    // number of non visited points.
    n = sizeOf(V);
end
C = sampleSeedTemplateFromImage(S, I);
for i ← 1 to n do
    p_i = sampleRandomPoint(V);
    removePointFromSet(p_i,V);
    t_i = sampleTemplateFromImage(p_i, I);
    addTemplateToSet(t_i, T);
end
forall the t ∈ T do
    s = calculateSimilarityScore(t_i, C);
    setSimilarityScore(s, t_i, M, p_i);
end
```

4.2 Applications for Object-Based Attention

In neuroscience and psychology object-based attention explains the behavior of neural responses when an object is fixated. Visual stimuli are adjusted in favor of the particular object, which enhances the processing of the object's features. Here we present two applications which can benefit from our object-based attention method: 1.) Object Recognition and 2.) Visual Search Tasks

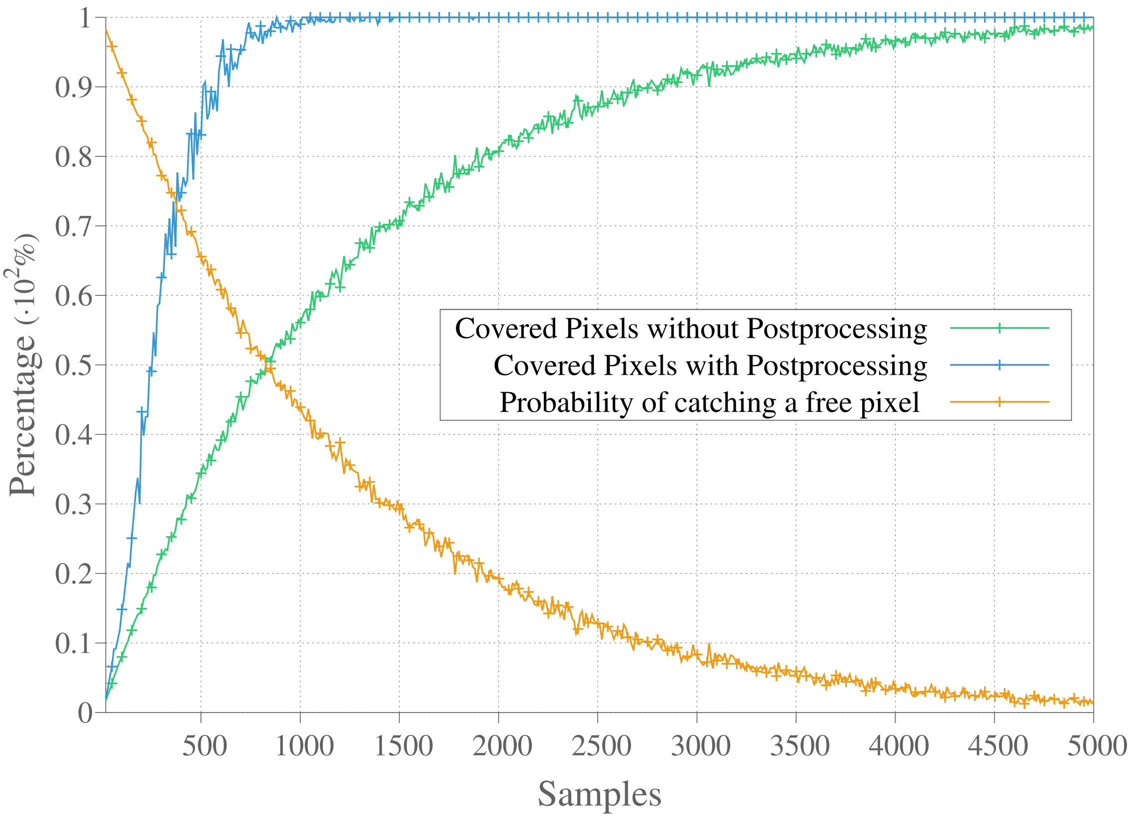


Figure 4.5 The effect of sampling on the number covered pixels and probability to pick a previously unvisited pixel using the sparse sampled template collation approach. Similar to the coupon collector's problem, the number of selected pixels which haven't been visited before significantly decreases with number of samples. Without post-processing it takes about 5000 samples for a 1200 pixel image to cover the whole image.

4.2.1 Object Recognition

Our Object-based attention approach enhances the representation of the attended object by biasing the object's features. This is achieved by segmenting the attended object from the surrounding, which is a crucial preprocessing step in object recognition. An image which contains multiple objects will produce ambivalent results in the classification because features are generated from all objects. Therefore we introduced object-based attention to the object recognition process.

More explicitly we can say that the object recognition process benefits from the object-based attentional approach for two reasons:

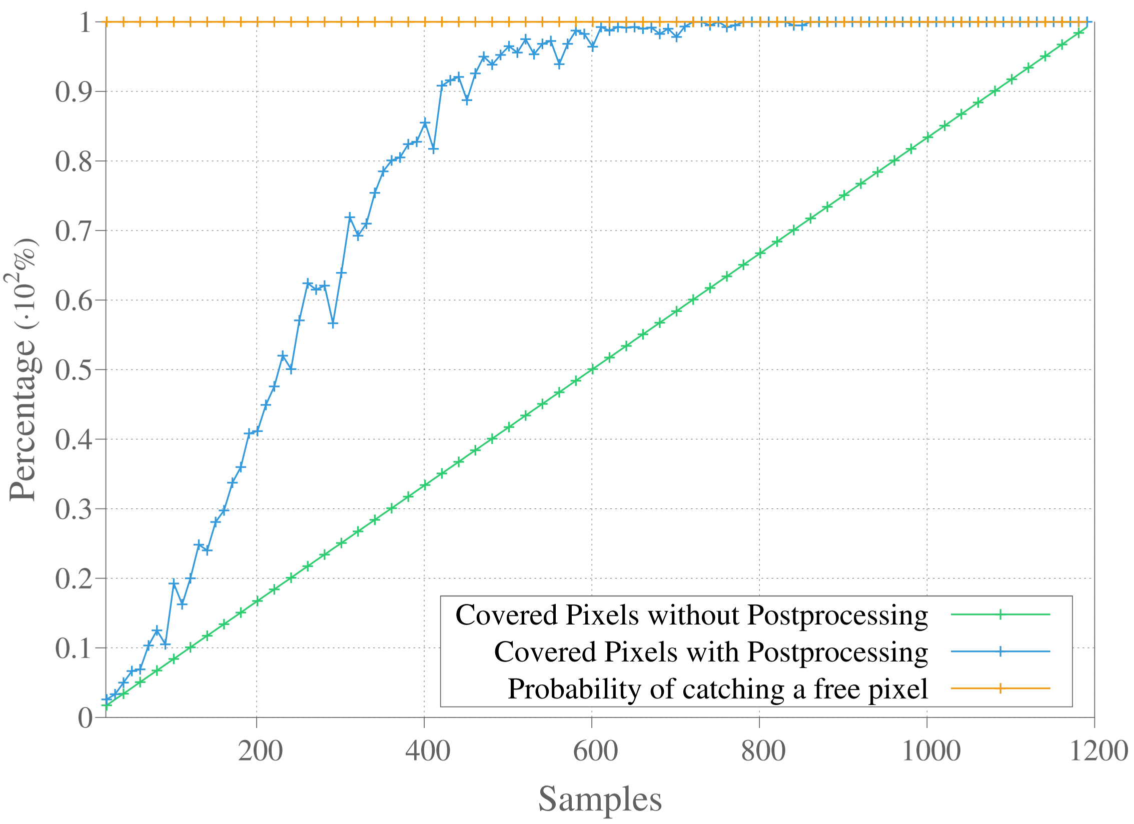


Figure 4.6 The effect of sampling on the number covered pixels and probability to pick a previously unvisited pixel using the efficient sparse sampled template collation approach. The number of new selected pixels which haven't been visited before a linear in the number of samples. No redundant computational step needs to be performed. Without postprocessing the samples needed to cover the whole image are identical to the number of pixels in the image.

1. It provides a segmentation of the fixated object from the surrounding areas which are likely to contain objects that interfere with the classification performance. This is especially useful in cluttered scenes, and essential for interacting and reasoning about the environment.

2. It drastically reduces the region of interest and therefore the area where the templates are sampled from. Subsequently less templates are needed to that cover the area of the object, which accounts for a faster processing speed. The sampling process in the object recognition step (see chapter 5) is particularly computationally intensive as the feature vector is generated calculating the response of every template in the dictionary to every newly sampled template, which results in a

complexity of $O(n \times m)$, with n being the number of sampled templates and m being the number of templates in the dictionary.

We tested our object-based attention approach on various objects (see figure 4.8). The tests show good results also with cluttered objects. The Object Recognition greatly benefits from this object segmentation step. Before, it was not possible to classify an image with two known objects of different classes in it. Our approach now enables a distinct classification of objects in the same image. Additionally the probabilistic classification over time function (see chapter 5) converges faster, because templates are sampled only from the object and therefore the classifier outputs a higher probability of the object's class.

We evaluate our approach by measuring the probabilistic responses of the classification with and without object-based attention. The results in figure 4.7 show, that the classification with our approach is more accurate and consistent compared to the previous approach. The probability estimates have less variance and are around 97%, whereas without OBA the results show higher fluctuation and significantly less accuracy with around 70%. This also benefits the probabilistic summation over time approach - the believe system achieves 100% almost immediately, without object-based attention it takes three times as long. The old approach was not able to distinguish between objects, whereas the new approach showed no difference in classification performance to single object images.

4.2.2 Visual Search

We modified Sampled Template Collation for object-based attention so it can be used for visual search tasks. Visual search is an active perceptual task involving attentional mechanisms. The environment is scanned for a particular object using the object's visual features. Many visual search model are biologically-motivated and based on the Feature Integration Theory explained in chapter 2. Visual Search provides clues of the position of the searched object, which are then verified in active saccades on the particular positions. We are not constantly aware of every object in our visual view and only objects we actually attend to are recognized in a higher cognitive fashion.

This approach is highly efficient from a computational perspective:

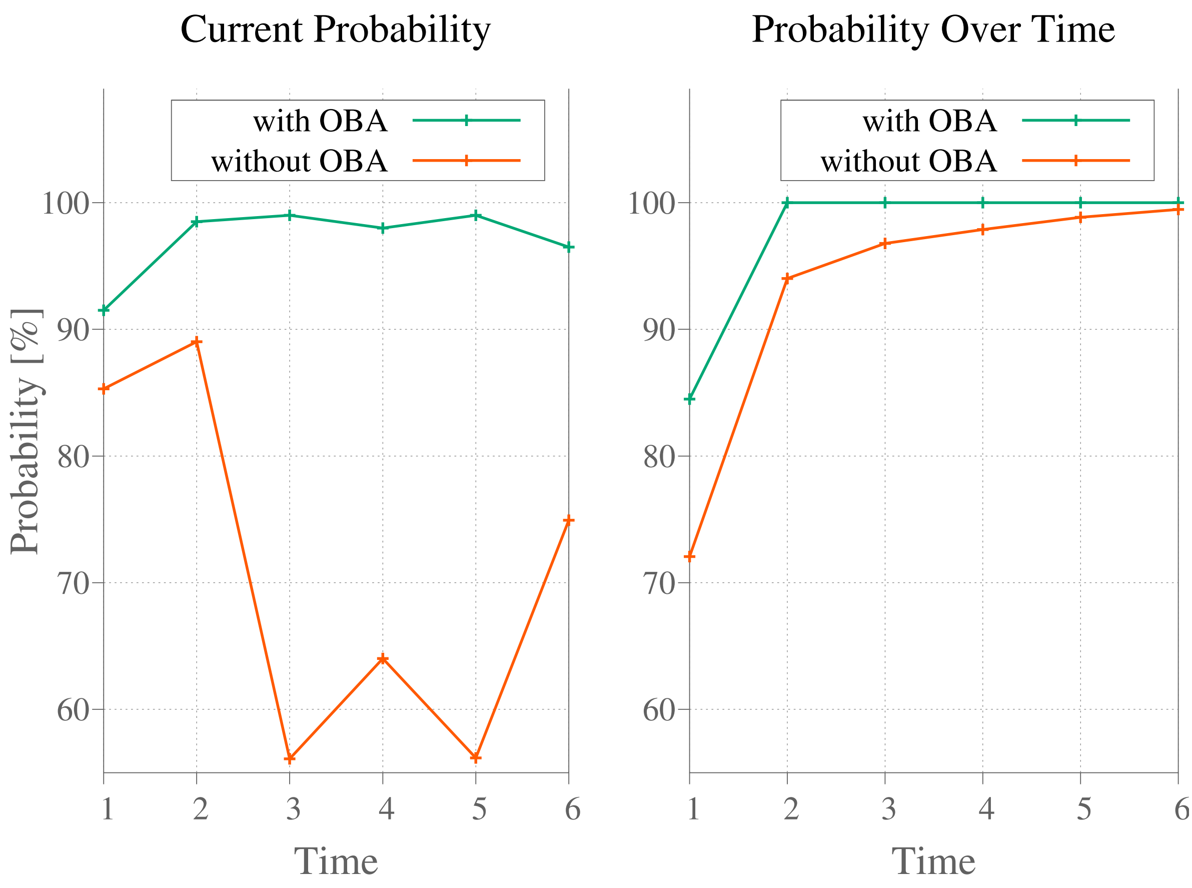
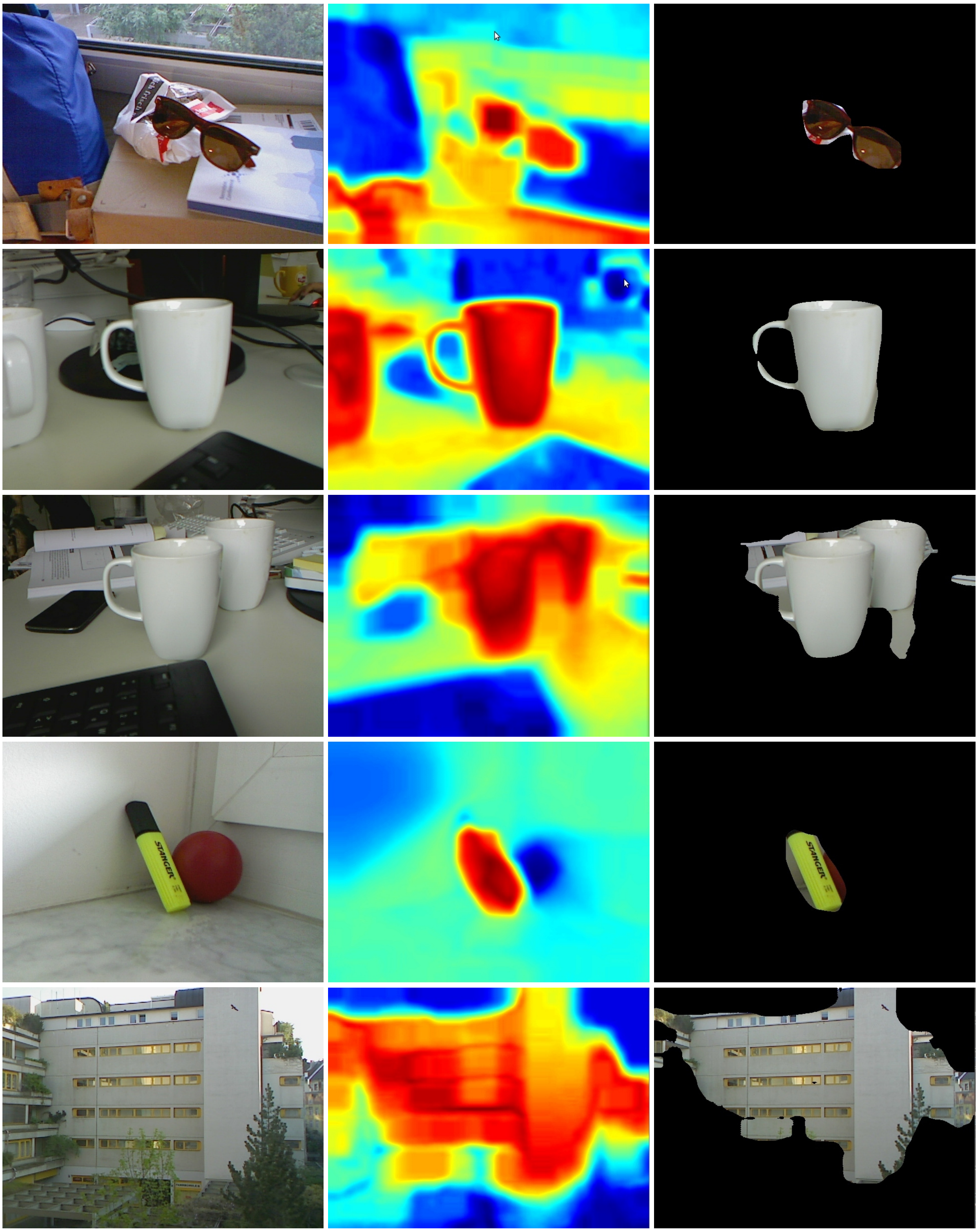


Figure 4.7 Classification results with and without Object-based Attention (OBA) for a test case with two objects in the image. The different images were acquired over multiple time steps at different view angles on the objects. The results show that the object recognition noticeably benefits from the OBA approach. The probability estimates are much preciser with OBA (green line) than without (orange one).

1. The information needed for visual search, which are the features in the visual field, is already present due to the feature processing for visual attention. This information is then biased in favor of features that are similar to the object we are looking for.
2. It drastically reduces the region of interest for higher cognitive processes like object recognition. First the image is segmented for possible candidates of the object and only then verified by computational more complex functions.



(a) Input Image (b) Object Heat Map (c) Intersection

Figure 4.8 Object-based Attention using Sampled Template Collation. The center of the input image (a) is used as template seed to create the object-based attention heatmap (b). Column c shows the result of intersecting heat map and input image.

We slightly modified our Sampled Template Collation approach to be able to bias those templates which might be sampled from the object (see algorithm 7).

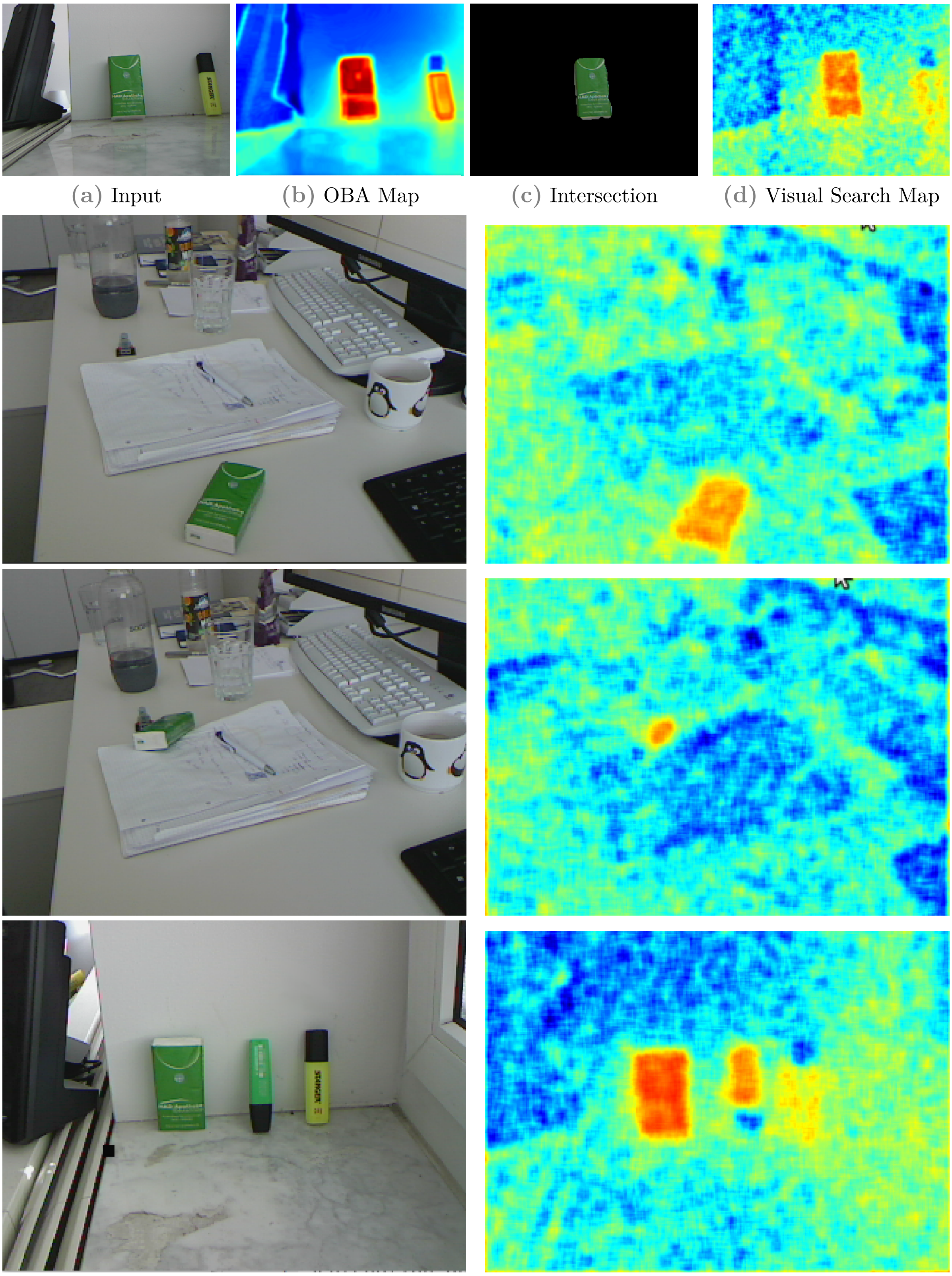
First we generate an object-based attention map and intersect the result with the input image. The resulting area of the segmented object is then used as a region of interest for the initial sampling process for generating a set of object templates (see figure 4.9 (A) - (C)). The sampled object templates now act as seed templates. For every sampled template in a new image, those object templates are used to calculate a similarity score. This score is then again back projected to the origin of the image template to generate the visual search map. Figure 4.9 shows the initial procedure (first row) and some results of the process. Figure 4.9 (D) shows the initial visual search map generated with the templates sampled from the intersection in (C).

Algorithm 7: Sampled Template Collation for Search Tasks

Data: Image I; Set of templates T; Salient Point S; Object-Based Attention Map M; Object Templates O; Visual Search Map V
Result: Object-Based Attention Map M

gaussianBlur(I);
convertToLabColorSpace(I);
M = createObjectBasedAttentionMap(I, S); O = sampleObjectTemplates(I, M);
T = sampleImageTempalts(I); **forall the** $t_k \in T$ **do**
  forall the $o_i \in O$ **do**
    s = calculateSimilarityScore(t_i, o_i);
    setSimilarityScore(s, V);
  end
end

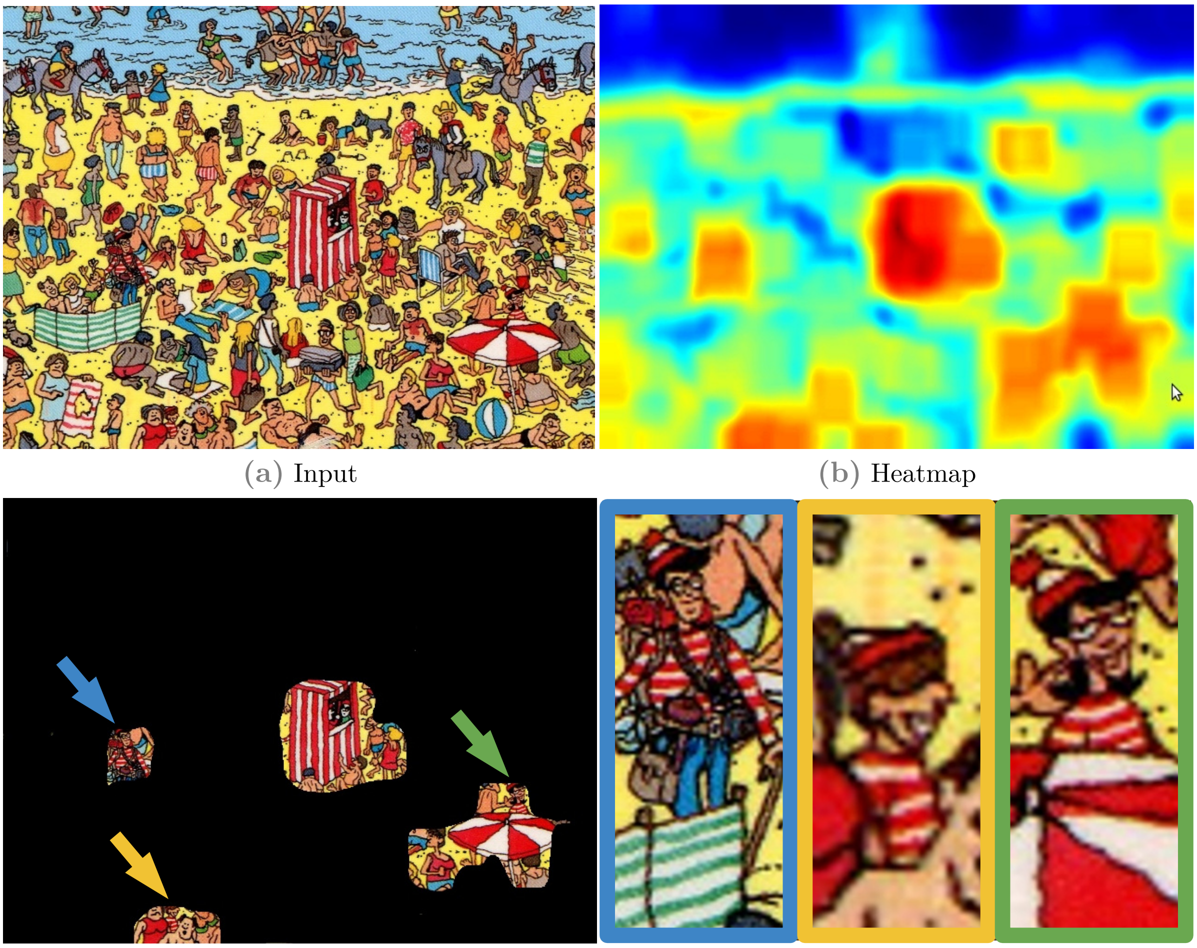
Figure 6.3 shows visual search results for Waldo. Here only one seed template is sampled from the middle of the image, which has a similar pattern as Waldo itself. This seed template is used to calculate a visual search map. The result contains three candidates, one of them is Waldo, the other two look very similar to Waldo.



(a) Input (b) OBA Map (c) Intersection (d) Visual Search Map

(e) Input (f) Visual Search Map

Figure 4.9 Visual Search with Object-based Attention using Sampled Template Collation. The first row pictures the initial step, where object templates are sampled from the green packaging from the created OBA intersection. The three following rows show examples of the visual search maps.



(a) Input (b) Heatmap

(c) Intersection (d) Candidates

Figure 4.10 Where's Waldo? A seed template is taken from the middle of the image (a), which has a similar pattern as Waldo itself. The heatmap (b) shows the highest responses for this pattern. Intersected with the input (c) the candidates can be extracted (d). [Image from http://whereswaldo.com/]

4.3 Summary

In this chapter we presented Object-based Attention using Sampled Template Collation. The method we developed is based on our sampled template collation model for visual attention.

We evaluated our approach and showed that it is online scalable, has low computational complexity and improves object recognition results due to the segmentation. We evaluated the dense and sparse sampled approach in regard to computational complexity and consistency and showed that sparse sampled template collation can achieve the same results with faster processing than the dense method. We also introduced efficient sparse sampled template collation which doesn't suffer the coupon collector's problem. We presented useful applications for object segmentation, visual search and object recognition.

In the next chapter we introduce our approach to object recognition. It is based on a simplified computational model of information processing in the visual cortex.

Chapter 5

ENHANCING A COMPUTATIONAL MODEL FOR OBJECT RECOGNITION ModHMAX

This chapter will introduce the object recognition system developed in this thesis. It consists of ModHMAX, a for time-crucial applications enhanced modification of HMAX and the concept of temporal reasoning, which introduces time to static recognition models and presents a more realistic approach to biologically-inspired object recognition. The chapter starts with a discussion about the use of 3D information in object recognition.

5.1 3D or not 3D?

In the last couple of years 3D sensing camera systems have gained in popularity due to inexpensive infrared emitting RGB-D cameras like the Kinect. Using depth for object recognition can have advantages over normal 2D approaches, especially because the depth information can be used to segment the environment by using the position information of spatially separated objects. Using depth information can additionally avoid misclassification due to texture or lightning.

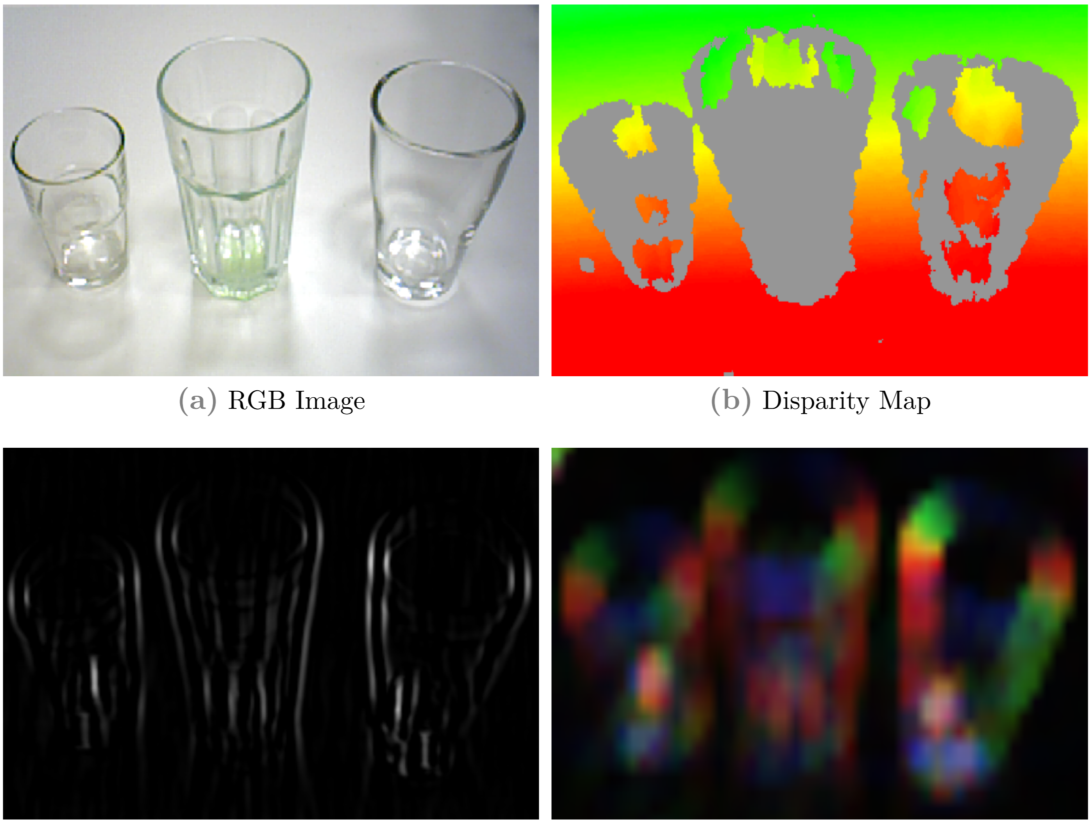
The methods to acquire depth information can be separated into two categories: Active and passive. Passive systems use two cameras, where position and distortion are known after calibration. The depth can then be inferred by point correspondences. Active systems use a light emitting source like infrared or laser and - in the first case - project known patterns onto a surface and therefore calculate the depth information. Systems with laser measure the time of flight the light took to the object and back and infer the distance. These so called Light Detection And Ranging systems (Lidar) can have a much higher range compared to the Kinect, but are also much more expensive.

One major problem with those systems is that light passes through transparent objects, returning no depth information at all. This is especially an issue with regard to household robotics, where a lot of the objects are made of glass or transparent plastic. To visualize this issue, figure 5.1 shows a disparity map of a scene containing glasses. The color gray in the disparity map means, that there is no depth information at all about that region and it is therefore not possible to create features for learning and classifying objects. Using 2D RGB information instead, we are able to create features like the one generated by our system.

The drawback with transparent objects and the fact, that depth information is not essential for humans to recognize objects are reasons to investigate 2D object recognition rather than 3D. Recent success stories in object recognition performance using convolutional neural networks on two dimensional data support this assumption.

5.2 The ModHMAX Computational Model

Our object recognition system is strongly based on HMAX, as it allows



(a) RGB Image (b) Disparity Map

(c) S1 feature map (d) C1 feature map

Figure 5.1 Comparison 2D and 3D. There is no depth information available using an active light emitting sensor (b) and therefore we have no information about the object. It is not possible to create features using only the disparity map. By using the 2D RGB image instead, it is possible to create features like the ones used in our system: Figure (c) S1 and (d) C1.

1. for a quick training with little training data,
2. the computational advantages for processing and classification, which is a crucial requirement in robotics,
3. its possibilities for improvements and modifications and
4. it proofed superior to state-of-the-art object recognition systems like SIFT (see section 2.3.2).

We developed and evaluated different methods and modifications in the architecture for improving the standard HMAX model to be applicable in real-world scenarios in terms of speed, object recognition performance and classification over time. Figure 5.3 shows the architecture of our system with the corresponding layer in the original HMAX model and the related area in the visual cortex. Figure 5.2 shows a map of the locations of the corresponding brain areas with the processing latencies for rapid scene classification.

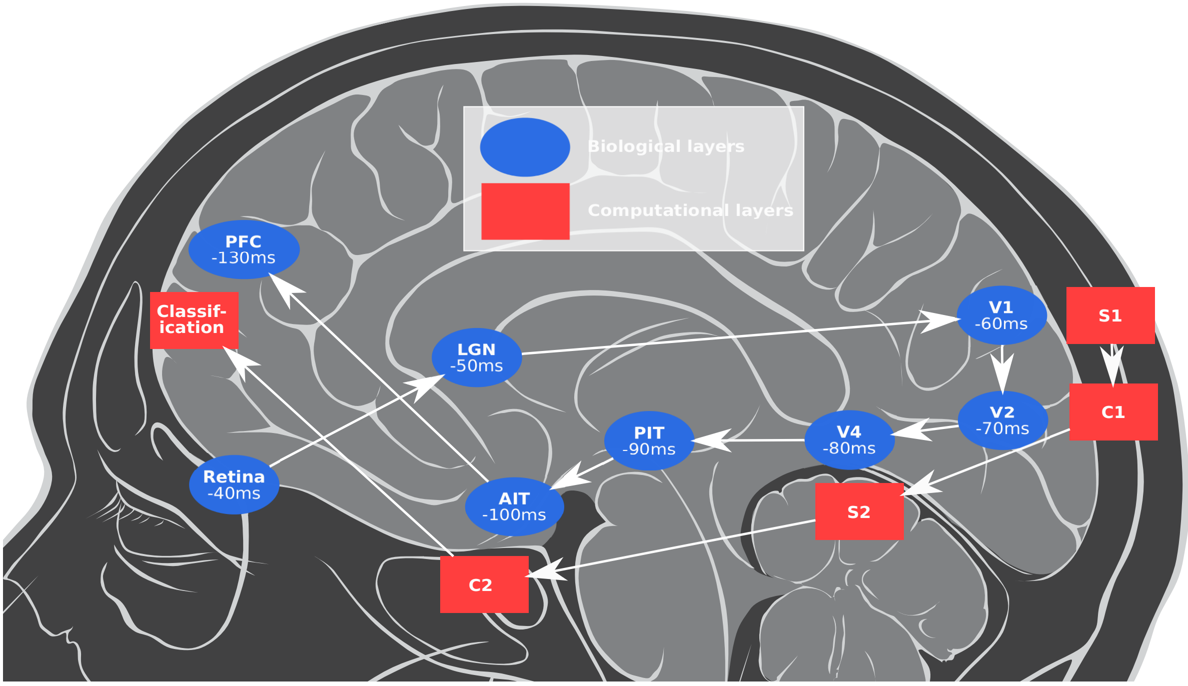


Figure 5.2 Processing latencies for visual stimuli in the brain. (Adapted from [Thorpe and Fabre-Thorpe, 2001]).

5.2.1 Enhancements and Modifications in S1 and C1

The first layer is based on a representation of simple cells in V1 which react to oriented edges and bars in their receptive field. The response of these cells is quite similar to Gabor filters with specific parameters according their tuning of orientation and frequency; The Gabor filters are created using the function

$$G(x', y') = exp\left(-\frac{x'^2 + y'^2\gamma^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \psi\right) \tag{5.1}$$

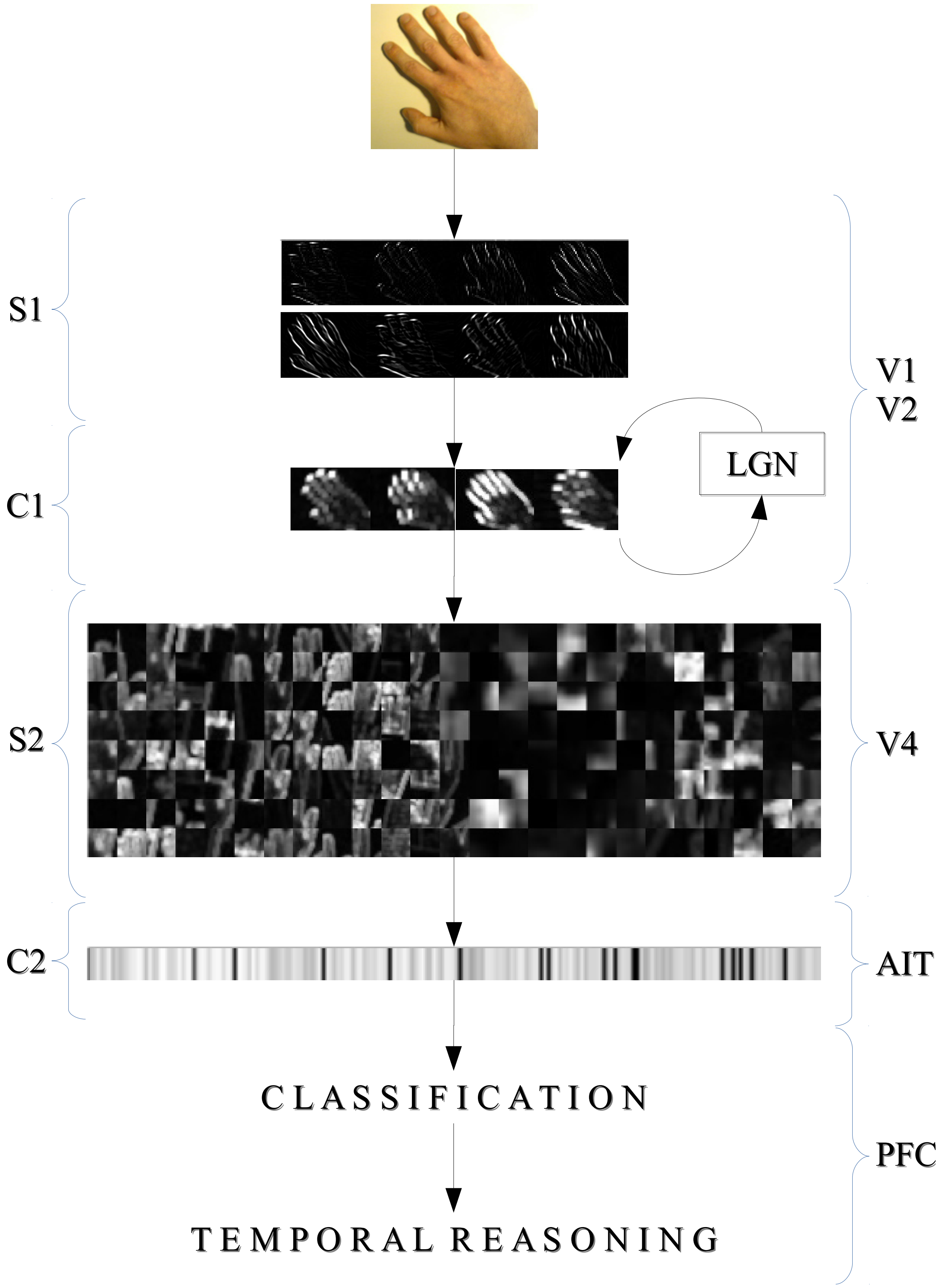


Figure 5.3 Functional Overview of our object recognition architecture. The left column indicates which response corresponds to which layer in the HMAX model. The right column gives a rough idea of the corresponding areas in the brain.

with

$$x' = x\cos\theta + y\sin\theta \tag{5.2}$$

and

$$y' = -x\sin\theta + y\cos\theta \tag{5.3}$$

where θ controls the orientation of the filter, ψ the phase offset, σ the variance of the Gaussian, γ the spatial aspect ratio and λ represents the wavelength of the sine function.

The original HMAX model creates a Gabor Filter Bank of 64 filters for convolution with the input image to create a representation of the S1 receptive field response. For biological plausibility and comparability we apply the same parameters as in Serre et. al. [Serre et al., 2007b], which resemble the response of the actual V1 parafoveal simple cells in the visual cortex; corresponding to neurophysiological data in [De Valois et al., 1982].

The second layer represents the complex cells in the visual cortex. They have a much larger receptive field than simple cells and add some degree of spatial invariance and shift tolerance to the system. They gain input from two S1 filter outputs of same scale band and same orientation. Their functionality can be described as a max pooling operation or a moving maximum over two filter outputs of S1; They keep only the maximum value of two neighbored (of same band) responses of the previous S1 layer within a sliding window.

An example: We have two windows W_a and W_b of S1 filter outputs a respectively b over the same area $[(x_0, y_0); (x_n, y_n)]$ with n being the window size. The C1 response at position (x_0, y_0) would be $max(W_a, W_b)$, with max being the maximal occurring value in W_a and W_b. The specific parameters and S1 neighbors (bands) are shown in table 5.1.

A single response r can be described as

$$r = max(W_{a_{x,y}}, W_{b_{x,y}}) \quad \forall x, y \in W_a, W_b \tag{5.4}$$

with W_a and W_b being one of the sliding windows sampled from the previous S1 layer over the same position of two neighbored filter outputs a and b; and x and y being all pixels in the window.

5.2.1.1 Orientation-free Gabor Filter

Gabor filters have shown to provide a good estimate for the response of cortical simple cells and so they are used in all of the HMAX-like implementations. The model presented in [Serre et al., 2007b] uses four different orientations with different sizes and parameters resulting in 64 different filters.

Mutch and Lowe [Mutch and Lowe, 2008] use a slightly different approach by applying twelve different orientations but with a sparse representation to a pyramid-based model. Instead of a template with multiple layers - one for each orientation - the sparse representations reduces the layers to one by applying a per-pixel maximum operation over the different layers. This results in a one layered template with only the most dominant orientation (see figure 5.4). Computing only one layer compared to four, reduces computational complexity - however computing twelve orientations per Gabor filter would result in 192 different Gabor filters[1] and therefore 192 convolutions, which is computationally worse than the standard approach with 64 Gabor filters with four layers.

According to the idea of sparse representation trough the introduction of more orientations and reduction to one layer, we investigated if it is possible to integrate the functionality to an earlier layer in the HMAX model in order to reduce the computational complexity. Therefore we had a look at the Gabor filter convolution at stage S1.

The different orientations are supposed to contribute to the system's orientation invariance. However, those models create n-layered patches - with n being the number of different orientations. These patches are used for creating a feature vector for classification by applying a radial basis function, which calculates the norm of the difference of the n-dimensional patches. Consequently the result of the RBF function is quite different if the patches are rotated, which indicates, that orientation invariance is in fact very limited. In addition Mutch and Lowe's template reduction indicates that a simple

[1] Mutch and Lowe actually use one sized Gabor filters applied to an image pyramid, which results in the same complexity.

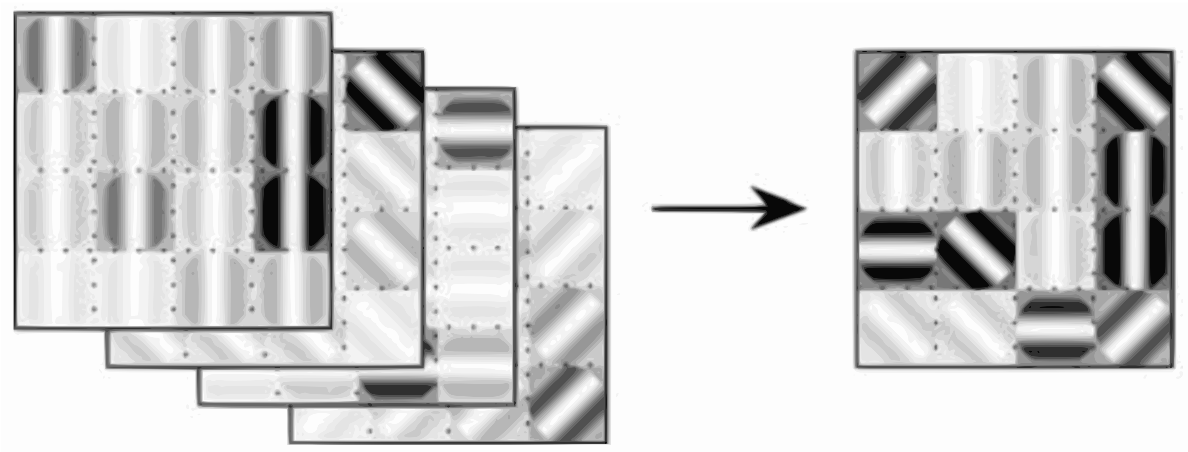


Figure 5.4 Dense representation (left) of a template compared to sparse representation (right) used in Mutch's and Lowe's model. (Image taken from [Mutch and Lowe, 2008])

combination of Gabor filters of different orientations could have the same effect while reducing the computational complexity.

Therefore we investigated, if Gabor filter of different orientations can be combined by creating an orientation-free Gabor filter using

$$G_{\lambda,\psi,\sigma,\gamma}(x,y) = exp\left(-\frac{x^2+y^2\gamma^2}{2\sigma^2}\right)\cos\left(2\pi\frac{\sqrt{x^2+y^2}}{\lambda}+\psi\right) \tag{5.5}$$

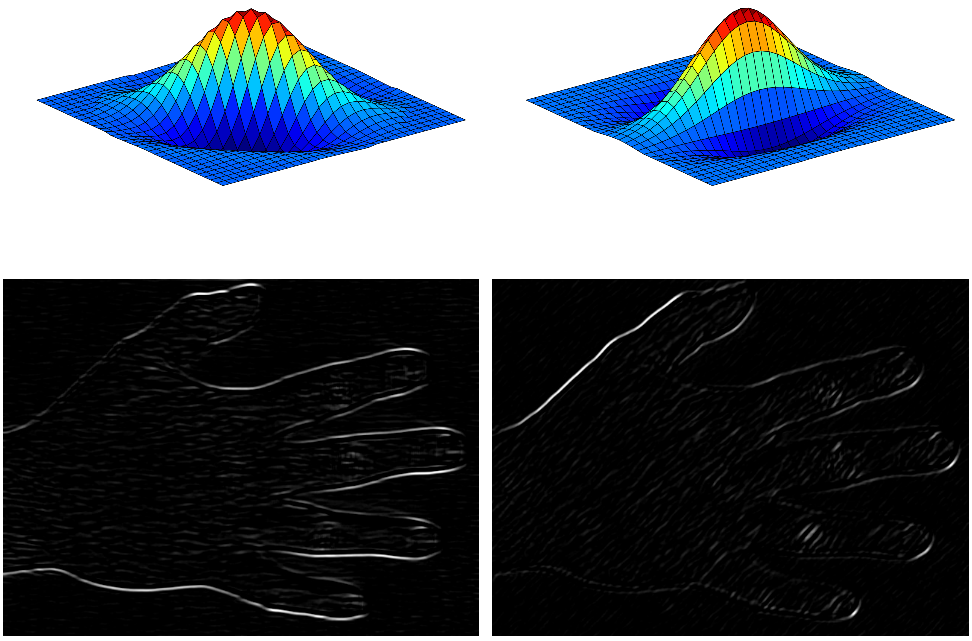
This approach creates a much finer representation of edges than ordinary Gabor filters, as all possible orientations are covered (see figure 5.6 and 5.5). Mutch et Lowe also argue in [Mutch and Lowe, 2008] that cells in the visual cortex have much finer gradiations of orientation than $\pi/4$.

In addition, an orientation-free Gabor filter reduces the computational cost of convolution from n orientations to one - in our case from 64 to 12. Another benefit of a orientation-free Gabor filter is that it is separable, which would make it computationally more effective. But in the HMAX model the filter is only defined within a circular area as it is more accurate to a simple cells' anatomy, which makes it non-separable. We tested non-circular Gabor filters against circular ones and got better defined edges using the original approach.

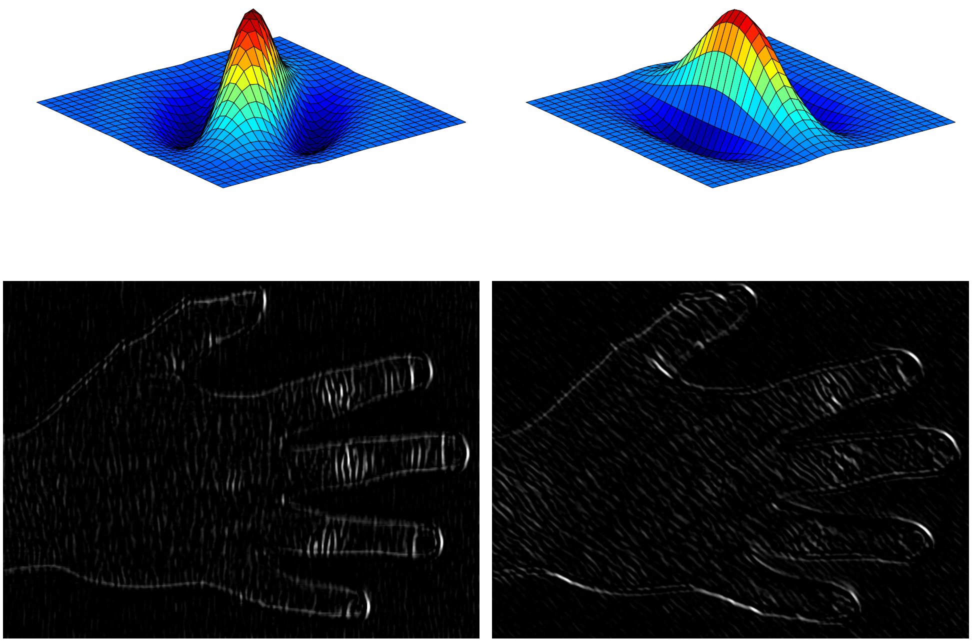
The features generated with the templates sampled from orientation-free Gabor filtered images showed however worse performance than the original approach (see figure 5.7). We believe that the features are too specific and complex to be used for classification. As this approach turned out to be insufficient, we investigated if the filter bank can be reduced while maintaining most of its descriptive features and classification accuracy.

5.2.1.2 Reducing the Filter Bank

The HMAX model has a set of fixed parameters for Gabor filters in the S1 layer and the MAX Pooling function in the C1 layer. The Gabor filter differ in size, amplitude and deviation. The C1 Filters in size, overlap (or stride). The applied parameters in both layers are organized in separated bands (see table 5.1). The response of a specific set of Gabor filters is processed using a specific grid size of the C1 layer. The reason mentioned in [Serre et al., 2007b] for choosing the specific parameters and bands is biologically motivated: the parameters seem to resemble the responses of Simple and Complex Cells. We are however more interested in performance than biologically plausibility, therefore we broke up the link between the S1 and C1 band and investigated how it affected the classification result. First we evaluated each band independently by creating features just by sampling from one single band at a time. The results in table 5.1 indicate that a larger filter size in S1 and C1 has a positive effect on the classification accuracy. In the second step we mixed the S1 band and C1 band and evaluated the effect on the test set. A larger Gabor filter size had a negative effect on the accuracy, while a smaller Gabor filter size improved the results. In the C1 Layer we measured an opposite effect, the larger the grid size of the max pooling filter, the better the results. In order to evaluate this effect, we added even larger MAX filter to the system, which are not present in the original HMAX model. A selection of the results of possible combinations are shown in table 5.2. It clearly indicates, that a combination of small Gabor filters and large MAX pooling filter are superior and outperform the other possible combinations. Therefore we chose the S1 filter band 2 with a filter size of $11x11$ and $13x13$ and the C1 filter band 9 with a size of 28.

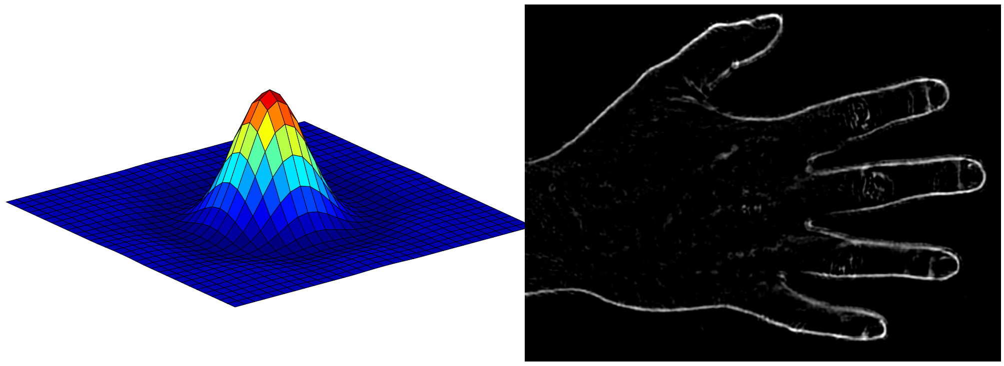


(a) 0° (b) 45°



(e) 90° (f) 135°

Figure 5.5 Gabor filters with four different orientations.



(a) Orientation-free Gabor filter (b) Response

Figure 5.6 Orientation-free Gabor Filter

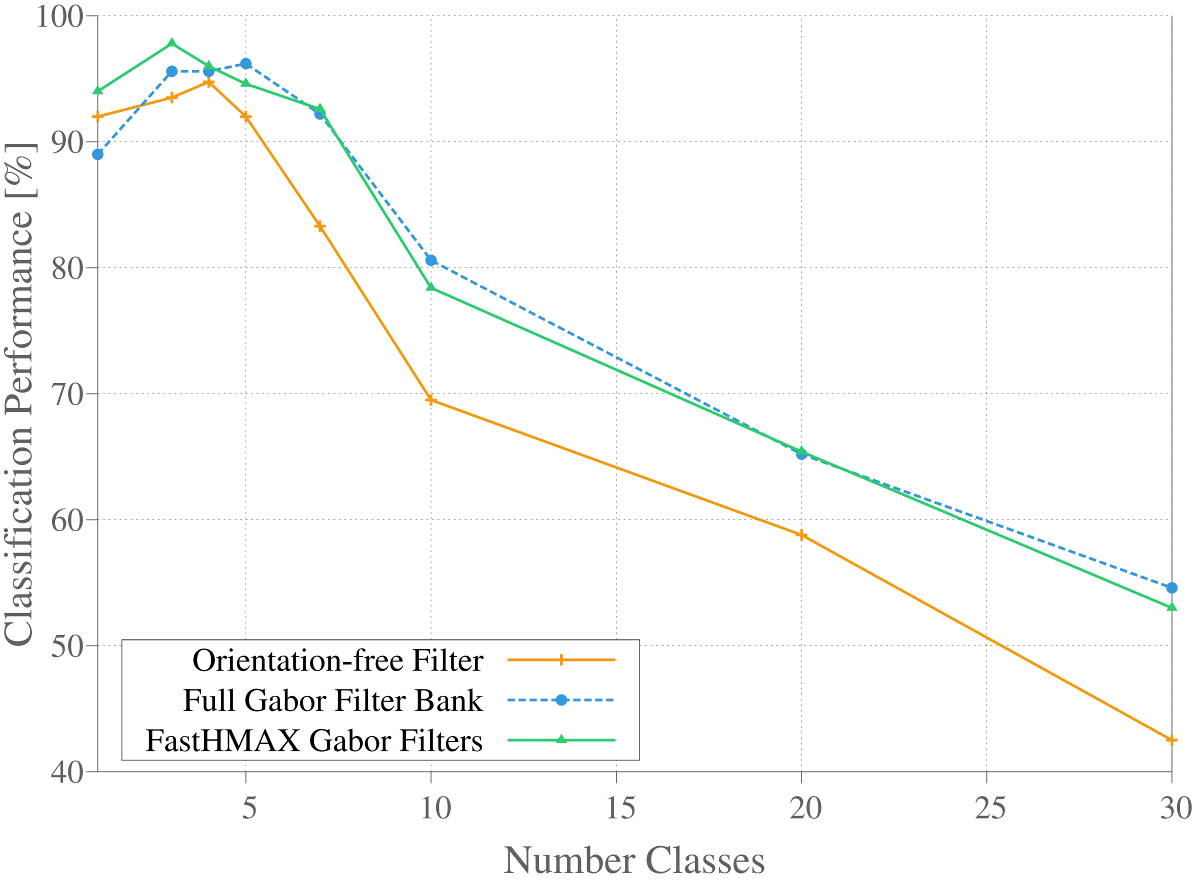


Figure 5.7 Classification Performance for Standard HMAX, Orientation-Free Gabor Filter and ModHMAX. [Dictionary Size 200; 800 Patches sampled.]

Table 5.1 Parameters applied in S1 and C1 by Serre in [Serre et al., 2007b]. On the right column, we evaluated the classification results for each single band individually. It indicates, that similar performance can be achieved using just a subset of the filter bank, because the difference in classification accuracy between a single band vs the whole filter bank is quite small.

	S1 Layer			*C1 Layer*			
	s	σ	λ	Grid Size	Overlap	Band	Accuracy
$\theta = 0°, 45°, 90°, 135°$	7×7	2.8	3.5	8×8	4	1	0.768
	9×9	3.6	4.6				
	11×11	4.5	5.6	10×10	5	2	0.774
	13×13	5.4	6.8				
	15×15	6.3	7.9	12×12	6	3	0.784
	17×17	7.3	9.1				
	19×19	8.25	10.3	14×14	7	4	0.780
	21×21	9.2	11.5				
	23×23	10.2	12.7	16×16	8	5	0.792
	25×25	11.3	14.1				
	27×27	12.3	15.4	18×18	9	6	0.804
	29×29	13.4	16.8				
	31×31	14.6	18.2	20×20	10	7	0.796
	33×33	15.8	19.7				
	35×35	17.0	21.2	22×22	11	8	0.788
	37×37	18.2	22.8				
					Original HMAX		0.806
					ModHMAX		0.842

Table 5.2 Classification results for different combinations of filter bands. Note that filter band 8 and 9 in C1 are not existent in the standard HMAX system. We added them to evaluate the effect of even larger MAX filter on the accuracy. Band 8 has size 26 and Band 9 has size 28.

Band S1	*Band C1*	*Accuracy*
1	7	0.808
1	8	0.812
1	9	0.808
2	1	0.784
2	6	0.814
2	7	0.821
2	8	0.840
2	9	0.842
3	6	0.798
3	7	0.814
6	1	0.754
6	7	0.808
7	1	0.738

5.2.1.3 Filter Factorization

Separable filter have a significant computational advantage over non-separable filters. A convolution in the spatial domain for a two-dimensional filter has the complexity $O(N * M * K^2)$, whereas a separable kernel has $O(N * M * K)$, with M and N being the number of columns and rows of the image and K the size of the kernel. Using singular value decomposition (SVD) we are able to factorize a Gabor filter into separable matrices. The SVD of the Gabor filter matrix takes the following form:

$$G = USV^T = \sum_{i=1}^{j} u_i s_i v_i^T \tag{5.6}$$

We can precalculate the separable filters and create the convolved image J from image I by using

$$J = \sum_{i=1}^{j} I * (u_i \sqrt{s_i}) + I * (v_i^T \sqrt{s_i}) \tag{5.7}$$

We achieve almost similar results for $j \geq 3$ compared to the original filter with an average error rate of $9.5 * 10^{-5}$ over the whole filter (see figure 5.8) - and still are faster by applying the separable filtering for $j = 3$ than using the non-separable filter.

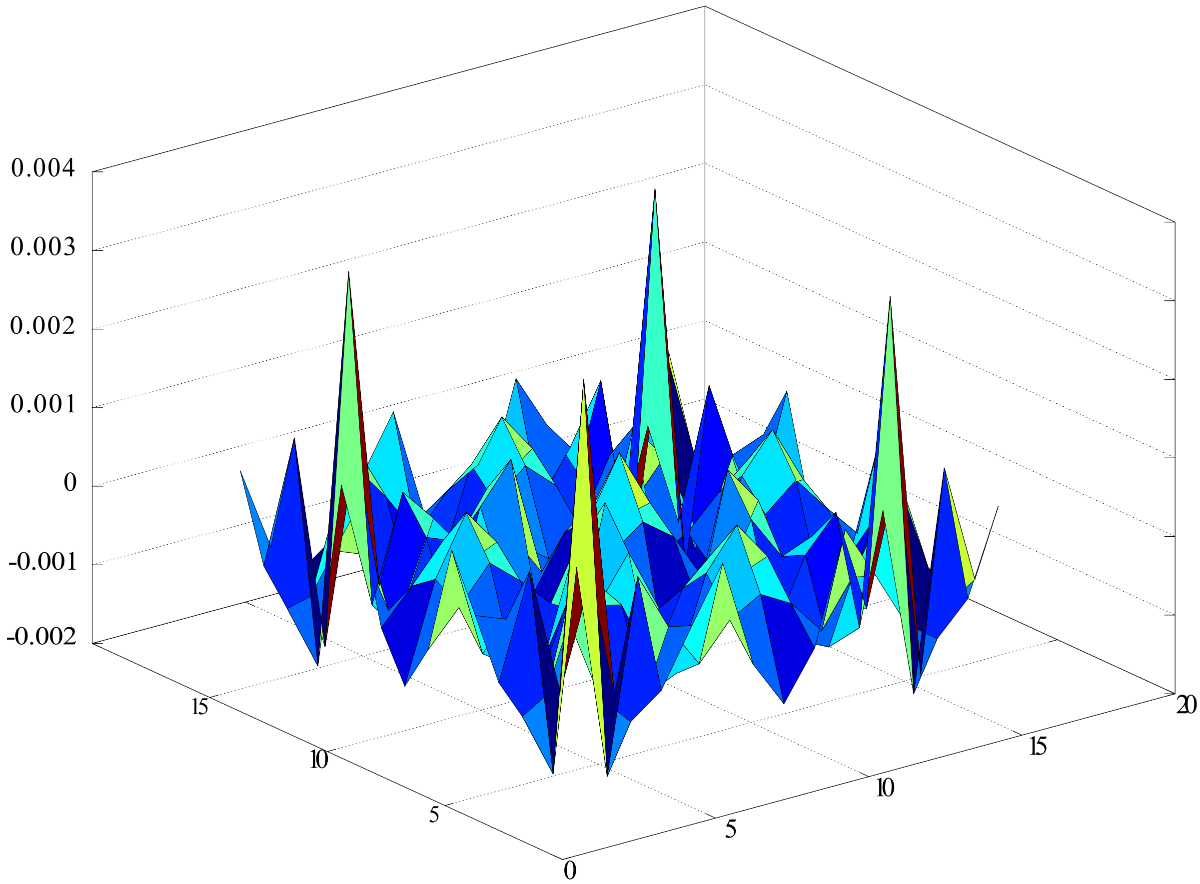


Figure 5.8 Error Distribution in a $20x20$ Gabor filter created with Singular Value Decomposition. The figure displays the difference between an original Gabor filter and one created using Singular Value Decomposition with the first three separated summations (see equation 5.6). The average error rate is very low ($9.5 * 10^{-5}$), so that this approximation can be used as an approximation for the Gabor filter.

5.2.2 Enhancements and Modifications in S2 and C2

The third and fourth layer of the HMAX model again mimic the functionality of simple and complex cells in the visual cortex. This behavior is represented with convolution and max pooling. Instead of Gabor filter like in S1, the used filters are templates randomly sampled during training from the C1 layer and collected in an universal dictionary.

Instead of template matching for n templates in the dictionary with k C1 feature maps, just a subset of all possible positions in C1 is randomly chosen and compared using a Radial Basis function.

Each templates has four channels, assembled by sampling from the C1 feature maps in the same band of different orientation $(0°, 45°, 90°, 135°)$ at the same randomly chosen position. Different template sizes are used, which slightly contributes to size invariance in the system. The templates are needed for two different cases: First for building a dictionary which is kept and used throughout training and classification, and second for calculating the response of new sampled templates to this dictionary which is used to create the feature vector.

The equation used to compute the distance between the template and a template in the dictionary is a radial basis function:

$$r_{i,k} = \exp(-\beta||T_i - D_k||_2^2) \tag{5.8}$$

with β being a weight constant for adjusting the amplitude of the response, T_i being a sampled template and D_k one of the templates in the dictionary.

Like in C1, the complex composite cells in the C2 layer perform a max operation over all the template responses across all scales, this creates an n dimensional feature vector, which can later be used for training a classifier.

For each template in the dictionary the maximum response for equation 5.8 is calculated using all the rbf responses of the templates of equal size. Using equation 5.8 this leads to

$$f_k = \arg\max_{T_i}(\exp(-\beta||T_i - D_k||_2^2)); \tag{5.9}$$

which builds the feature vector $F = \{f_0, f_1, \dots, f_d\}$ for all k in the dictionary D, with d being the length of the dictionary and T_i being a sampled template.

The operation removes all position and scale information resulting in global invariance. The whole response is a complex feature vector which can be used to train and test a classifier.

5.2.2.1 The Dictionary

The convolution and max pooling operation involving the dictionary is computationally expensive because the radial basis function is calculated between all randomly sampled templates T_i and all templates D_k in the dictionary. Note that the patch sets have different sizes {4,8,12,16}; so not all patch sets can be compared to another, which leaves a total number of RBF function calls of

$$n = |S| \times \sum_{m=0}^{|S|} |T_m| \times |D_m| \tag{5.10}$$

with S being the set of the different patch sizes, in our case $|S| = |\{4; 8; 12; 16\}| = 4$; T_m being the set of sampled templates with the patch size S_m and D_m being the set of templates with patch size S_m in the dictionary. Assuming for example $|T| = \{250; 250; 250; 250\}$ and $|D| = \{250; 250; 250; 250\}$ the number of rbf operations would be $4 \times 250^2 = 250.000$.

Reducing the complexity while at the same time keeping the classification performance at an acceptable level has received only little attention although it is crucial in order to meet the requirements for online classification. We investigated different possibilities to optimize the system for online processing by minimizing the size of the dictionary and the size of the sampled templates while maximizing the classification performance.

We investigated and evaluated different aspects of the dictionary. First we analyzed the changes in classification performance for different sizes of the dictionary. Figure 5.9

shows the results for a 10-class problem for a constant template sampling size of 800 and dictionary sizes of 20, 200, 2000 and 8000. At a dictionary size of about 200 we already get a significant improvement over smaller dictionaries whereas at a size above 2000 there are only little changes in the performance. While the orientation-free Gabor filter can't achieve equal results, the original HMAX and our ModHMAX modifications perform similar.

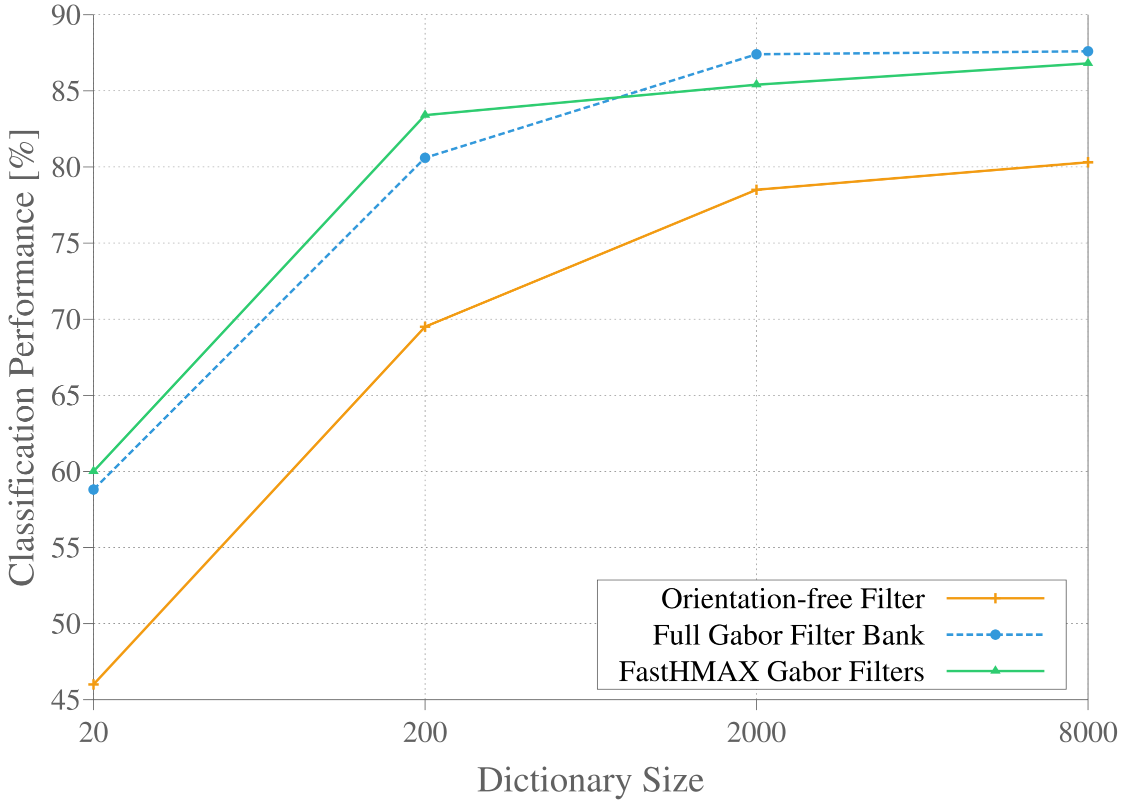


Figure 5.9 Classification Performance for Standard HMAX, Orientation-Free Gabor Filter and ModHMAX for different dictionary sizes. [10 classes; 800 Patches sampled in C1 layer.]

We further analyzed how the single features in the dictionary contribute to the classification performance. Therefore we evaluated different sizes of subsets of the dictionary with randomly picked features - Figure 5.10 shows the results of picking an optimal subset of the existing dictionary. In general this method improves the classification performance for smaller dictionaries, as important features are detected and kept in the subdictionary whereas bad features, that might even decrease the performance, are rejected. We evaluated the features using a statistical F-score, which is more significant than precision or recall, and applied the evaluation on a dictionary with 8000 features. ModHMAX in particular benefits from this approach and performs better than the original HMAX approach.

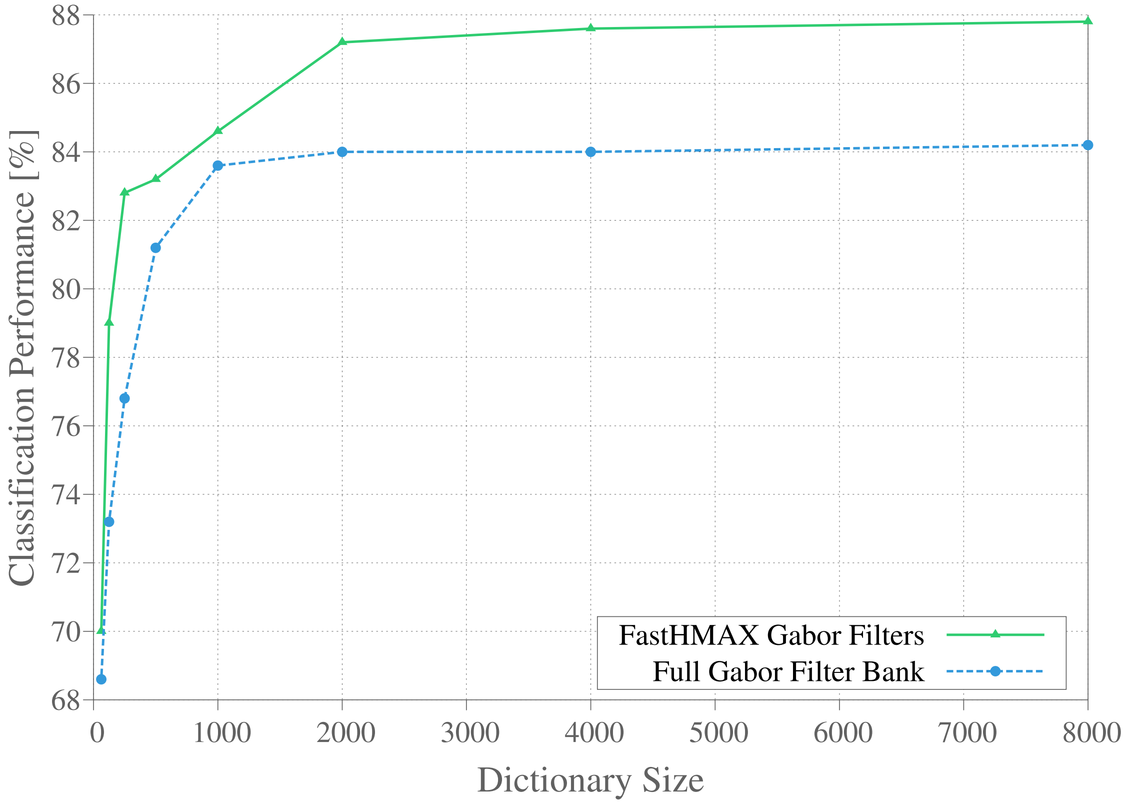


Figure 5.10 Classification Performance using optimal feature selection for the dictionary. [10 classes; 800 Patches sampled in C1 layer.]

We also investigated the influence of different template sizes in the dictionary. The applied sizes are 4x4,8x8,12x12 and 16x16. We evaluated the sizes by selecting only one template size for training and classification for different dictionary sizes and sampling rates. Figure 5.11 shows the results. Smaller templates seem to perform less well than templates with a larger size, as long as the dictionary and sampling rate is high enough. A small dictionary and a low sampling rate decrease the result for the largest template size - presumably because larger templates are more complex and characteristic only of a subset of objects.

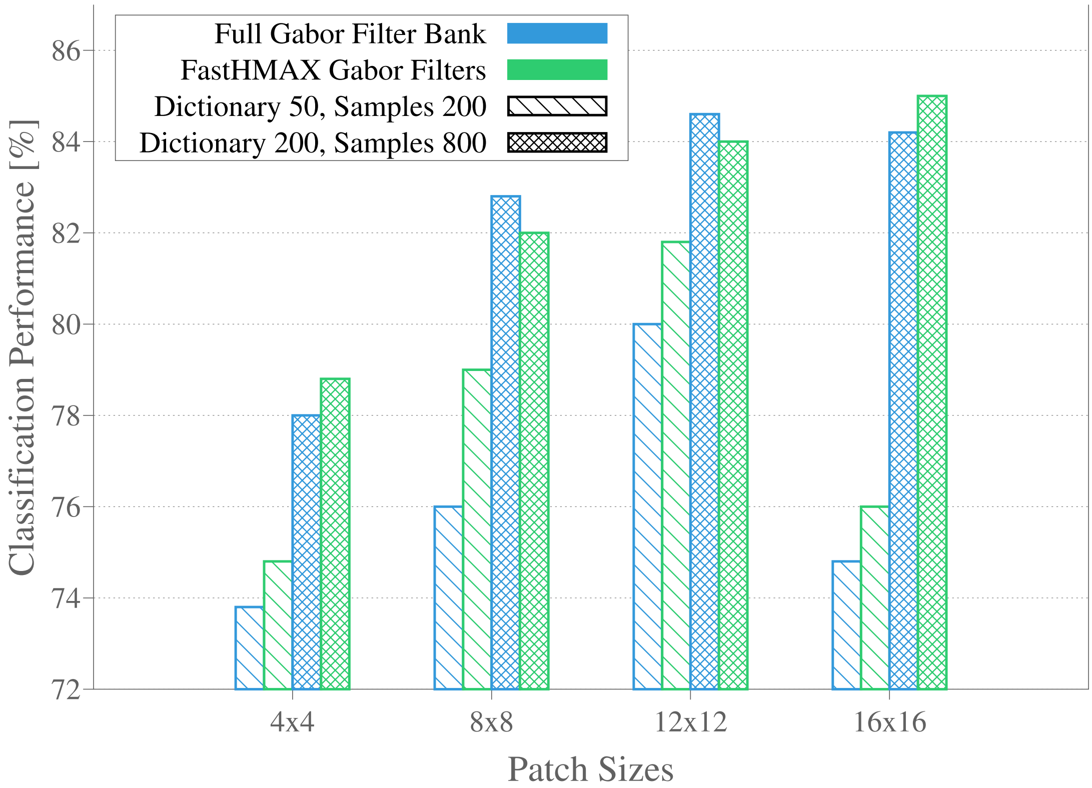


Figure 5.11 Classification Performance for Standard HMAX and ModHMAX using just one patch size. [10 classes; Dictionary Size 50 and 200 samples or dictionary size 200 and 800 samples]

We further investigated the effect of the sampling rate in the C1 feature maps on the classification performance. Figure 5.12 shows the results with a dictionary size of 200 and different sampling rates for a 10-class problem. At about 800 samples the performance reaches a maximum. ModHMAX outperforms the standard approach for a lower number of samples.

We tried to optimize the selection of templates in the dictionary in order to find a better representation of the objects. Therefore we applied k-means clustering to find centers in the sampled template sets by comparing the distance of intensities at each pixel position. The results however showed a decrease in classification performance (e.g. about 3% for the 10-class problem), this effect is consistent with the work described by [Serre et al., 2007b]. The k-means algorithm tries to minimize the distance of the templates to their assigned cluster center. We believe that by using k-means, small variations in

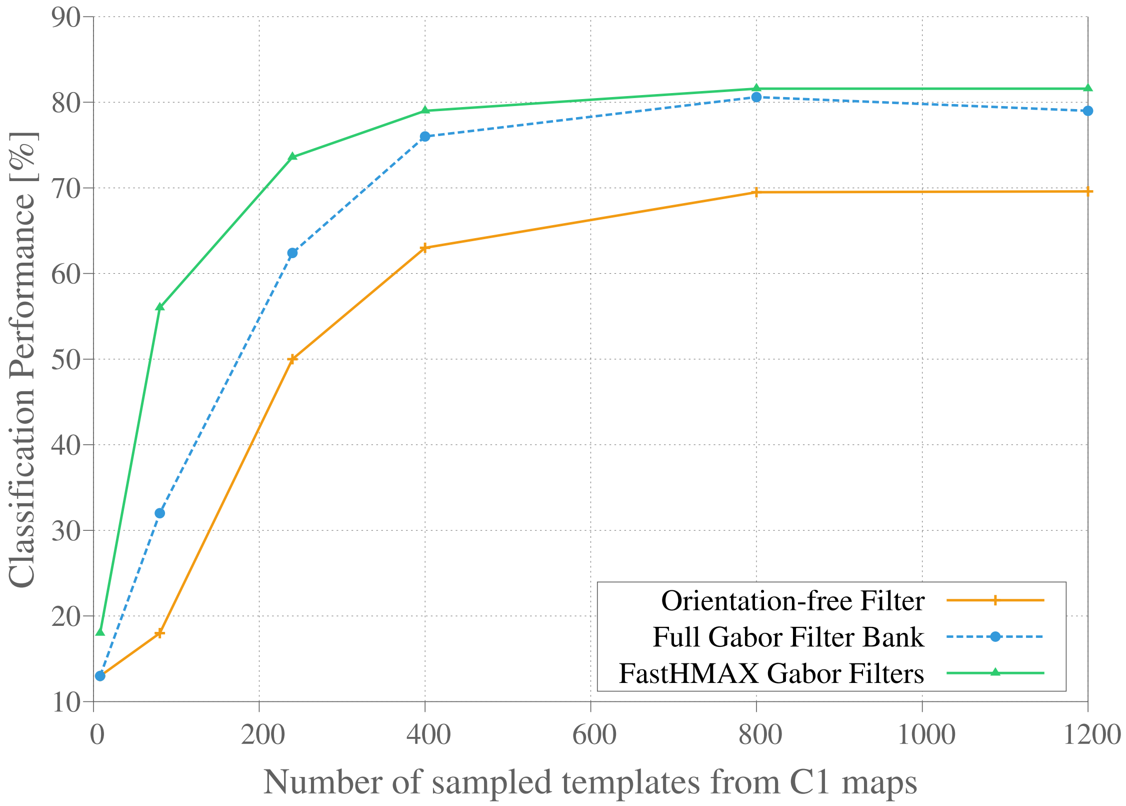


Figure 5.12 Classification Performance for Standard HMAX, Orientation-Free Gabor Filter and ModHMAX for different sampling rates. [10 classes; Dictionary Size 200]

the templates, which might be important for classification, are assigned to one common cluster and represented by only one center template. Additionally, the quantity of similar templates, which represent similar parts of an object is lost, whereas possible outliers might get represented with an own cluster center. By sampling randomly from a set of templates, these outliers are less likely to be selected.

In test scenarios which especially focus on computational performance using a small dictionary and low template size, the ModHMAX approach seems to outperform the original approach. A likely explanation is that the possibility to find a better template match with ModHMAX is more likely as the dictionary and the sampled templates are chosen from less feature maps as the original HMAX. For a larger dictionary and sampling size the original approach slightly outperforms ModHMAX.

5.2.2.2 Object Localization

In the standard HMAX implementation, the dictionary is created by randomly sampling templates as elements in the dictionary from the maps created in C1. This approach bears the risk to select a non-optimal set with over-represented and redundant features. Especially in image data sets, where image categories are presented in clutter for training and testing it is uncertain if the applied algorithm actually classifies the object itself or just the surroundings. The category car in the Caltech101 database is for example such a case: The actual object only takes a fraction of the image, whereas objects like trees or houses take up most of the space. Therefore it is uncertain, if the presented algorithms actually recognize the class car or mainly the background, as the templates are randomly selected over the whole image.

Instead of random selection, we wanted to analyze how well a dictionary can represent a single object class. To deal with this problem our method follows an approach, which is based on neural tuning. Cells in the brain selectively represent specific sensory patterns. We assign templates to specific object classes and each class is represented by an own sub-dictionary, which is created by keeping only templates that reoccur to a certain degree in all the training images. Hereby we want to achieve, that the created dictionary represents the actual object instead of it's surroundings. A car tire probably will appear in all images for example, however a tree might not, therefore patches containing the tree will most likely be filtered out.

After the sub-dictionaries are created, we apply an approach derived by lateral inhibition appearing in neural processing. For each template in a sub-dictionary we calculate the response of each template of each other sub-dictionary. If a template exists, which reacts above a certain threshold to templates in all sub-dictionaries, then these templates are completely removed. That way the sub-directories are more confined to their specific class.

Mathematically, we can describe the set of sub-dictionaries as a partition of dictionary D

$$D = \bigcup_{D_i \in D} D_i \tag{5.11}$$

with

$$D_i = \{x | \forall x \in D_i : \nexists y \in D_j, i \neq j : r(x, y) > \theta\} \tag{5.12}$$

with θ being a threshold of the response of the radial basis function r of Equation 5.8. Pseudo-Algorithm 8 displays how a sub-dictionary is created.

Algorithm 8: Create Object Specific Dictionary

```
Data: Sub-Dictionary D_i; Set of training images T; Set of patches C; Threshold θ
Create New Set Of patches(T_1, D_i);
forall the s > 1 do
    Create New Set Of patches(T_s, C);
    forall the k do
        forall the p do
            if f(D_{i_p}, C_k) < θ then
                delete(D_{i_p});
                break;
            end
        end
    end
end
```

One common method in computer vision to localize specific objects in an image is the simple sliding window approach, which is rather naive and inefficient, especially for time-crucial scenarios. The templates in the sub-directories are object-specific and therefore allow us to deduce the object location to a certain degree using the patches maximum response occurrences in the image. This approach requires no additional calculation, as the maximum responses are anyway needed to be calculated by the system in order to create the feature vector for the classifier. We create a saliency map by adding the maximum response values for each patch in the sub-dictionary to the location in the saliency map where the patch from the test image was sampled that created this highest response. Figure 5.13 shows some results of the object localization approach. We trained the dictionaries for a one-class problem using training images for the specific object as positive data and background images from the Caltech-101 data set as negative classes. Column B shows the salient regions calculated by backprojecting the maximum responses of the subdictionary to the sampled templates. Column C shows the salient regions of the subdictionary of the negative class. Note that there is no classification involved, the results are simply the maxima of the radial basis function values and therefore the backprojected feature vector values.

5.2.2.3 Integration of Entropy

There has been an ongoing debate about the feedback to the LGN[1] and its functional role is still not completely clear [Eivind Norheim and Einevoll, 2009; Briggs and Usrey, 2011; Jones et al., 2012]. The general consensus however seems to be, that there exists an influence to the cells in LGN generated by the feedback from V1.

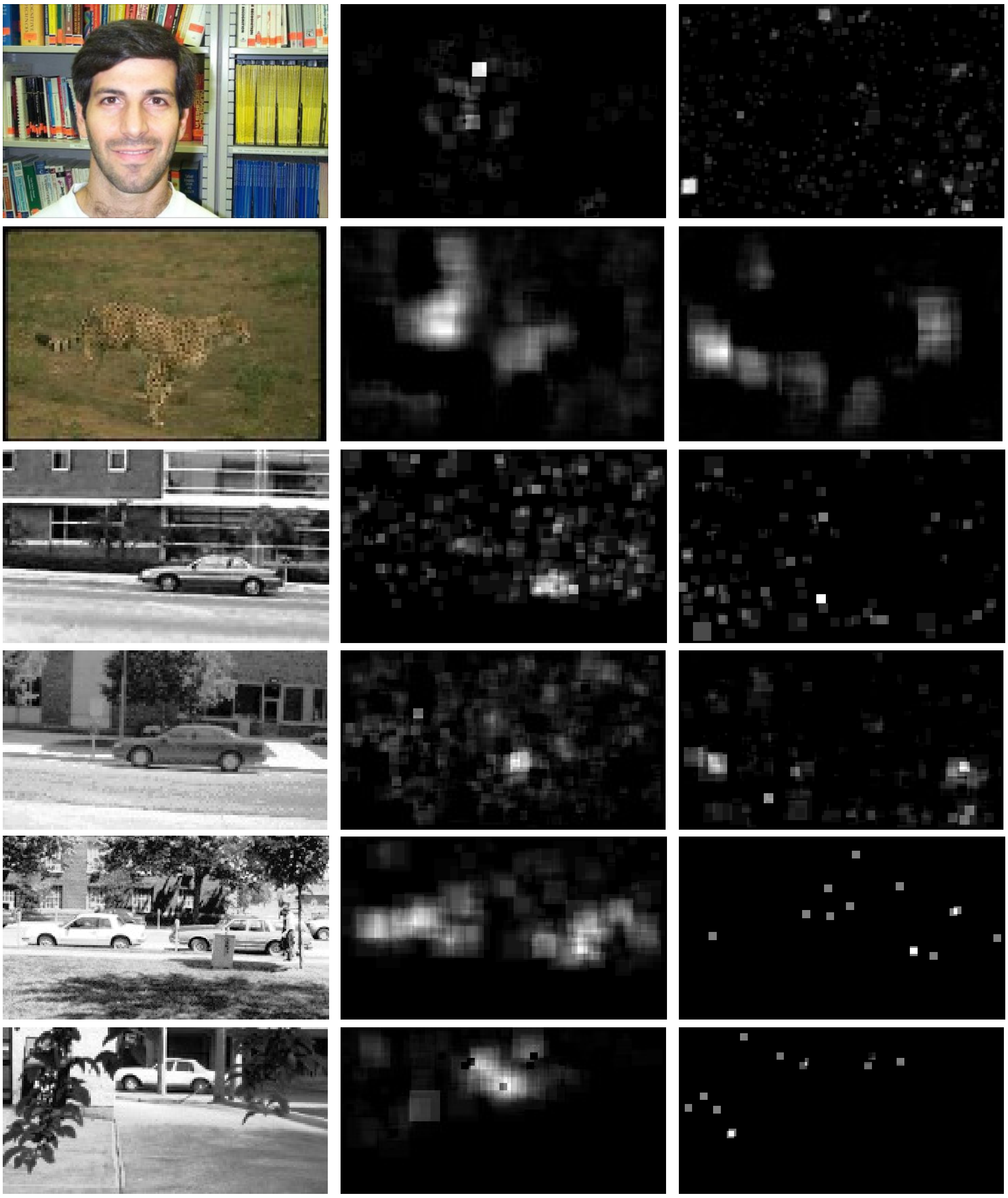
The standard HMAX model is a straight feed-forward hierarchy. This is an arguable simplification, as feed-back projections are an integral part of the visual system, although it might appear that those projections are counter productive to a low latency and fast processing. Here, we focus on the feedback from V1 to the LGN, due to the fact that cells in V1 provide an extensive feedback connection to LGN (about 30% of the synaptic input to LGN relay cells) [Sillito et al., 2006]. McClurkin et al. [McClurkin et al., 1994] state that the influence of feedback to the LGN enhances the information about the stimulus in the firing pattern. They cooled the visual cortex to reduce the feedback to the LGN and then applied Shannon's information measure to reveal that the average stimulus-information transmitted decreased. They conclude that the feedback increases the information that LGN neurons transmit in about all of the stimulus parameter they tested: pattern, luminance, and spatial and sequential contrast.

We integrated the information maximization and the feedback processing in the LGN in our object recognition architecture at the end of layer C1 and before the template matching in S2. The set of sampled templates are evaluated by the LGN node according to their information gain. Here we apply simple tools from information theory, and measure the self entropy for each randomly sampled template X by calculating

$$H(X) = -\sum_{m=1}^{M} p_m \log p_m \tag{5.13}$$

with p_m being the probability that a randomly chosen pixel in the template has brightness level m. $H(X)$ is large if the system has many equally likely states (high uncertainty);

[1]The lateral geniculate nucleus is located in the thalamus. It is the major target of the retinal ganglion cells. It receives inputs from both eyes and relays these messages to the primary visual cortex via the optic radiation



(a) Input image (b) Saliency map for object subdirectory (c) Saliency map for subdirectory of negative class

Figure 5.13 Object Localization. A saliency map of maximum responses to the object subdirectories. The map which belongs to the object in a) is shown in b); c) shows the response of a different object subdirectory. The first two images were taken from the Caltech101 database, the others were taken from the UIUC car dataset.

and therefore the template contains a higher information value. On the other hand, $H(X)$ is zero if and only if the system attains only a single state with $p = 1$. In that case the template contains no information. In our case, a template's entropy is gained by taking the intensity values of each position in the template into account. Pseudo-algorithm 9 describes how this step works: We sample a template, then evaluate its information value and depending on a threshold decide if we reject it or not. We do this until the size of the set of accepted templates matches a predefined sampling rate. To avoid an endless loop which can occur during online processing when the camera is for example pointed at a white wall, we set the threshold to reduce over time.

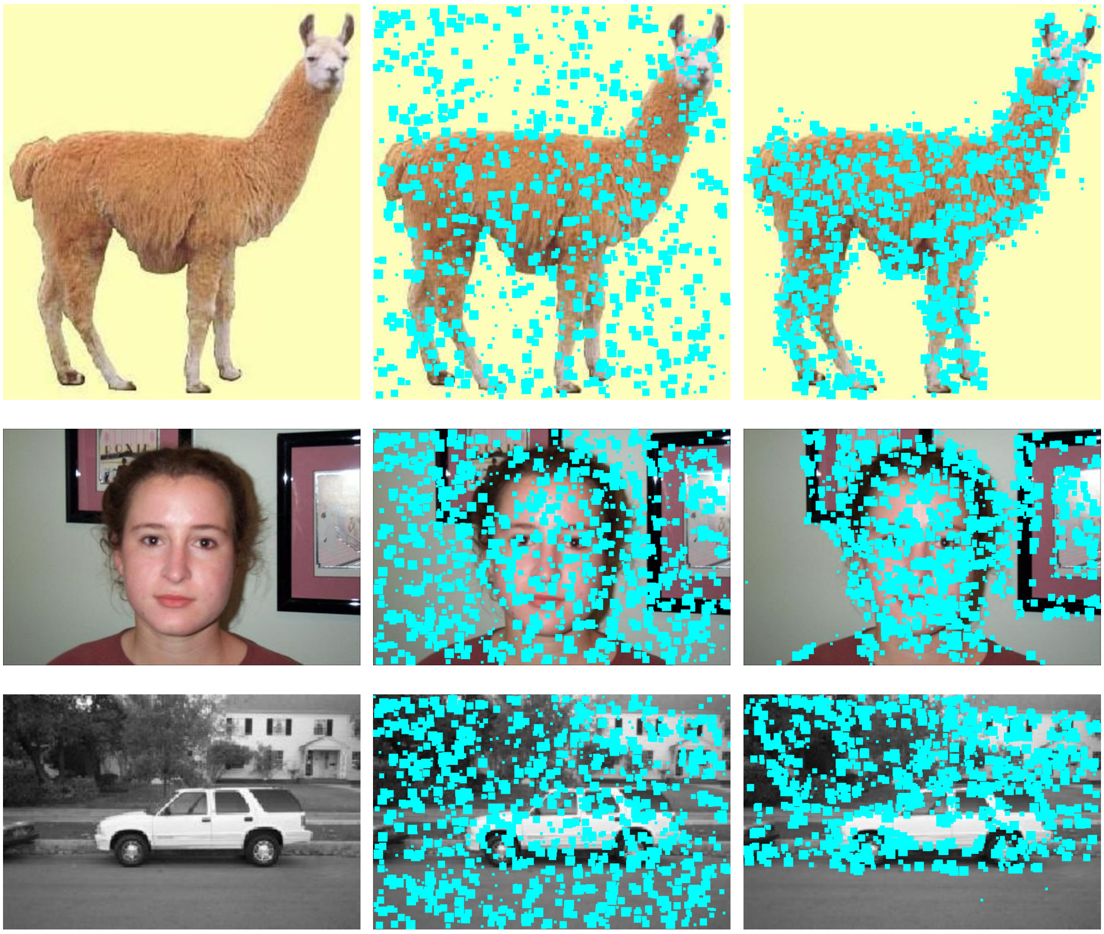
Algorithm 9: Entropy Integration

```
Data: Set of C1 feature maps M; Set of Templates T; Sampling Rate s; Threshold
      θ
while |T| < s do
    t = Sample new Template from M;
    h = Calculate Entropy(t);
    if h > θ then
        Add to Set of Templates(t, T);
    end
end
```

The LGN increases the information about stimulus pattern, luminance and spatial contrast [McClurkin et al., 1994]. We approximate this functionality by calculating the entropy of each template sampled randomly from an image. Templates with a low entropy are rejected and only templates with a higher entropy are kept. Figure 5.14 illustrates the impact of the LGN functionality integration to selectivity and sensitivity of templates on natural images. The blue dots represent sampled templates, those with a high entropy are shown in column c, in contrast column b shows the randomly sampled templates without the entropy step.

In order to further reduce the computation time of the system we tested two additional approaches to approximate the entropy in a template: 1. The standard deviation of the patch and 2. The difference of the maximum and minimum occurring intensity in the patch T:

$$H(X) \approx max(T) - min(T) \tag{5.14}$$



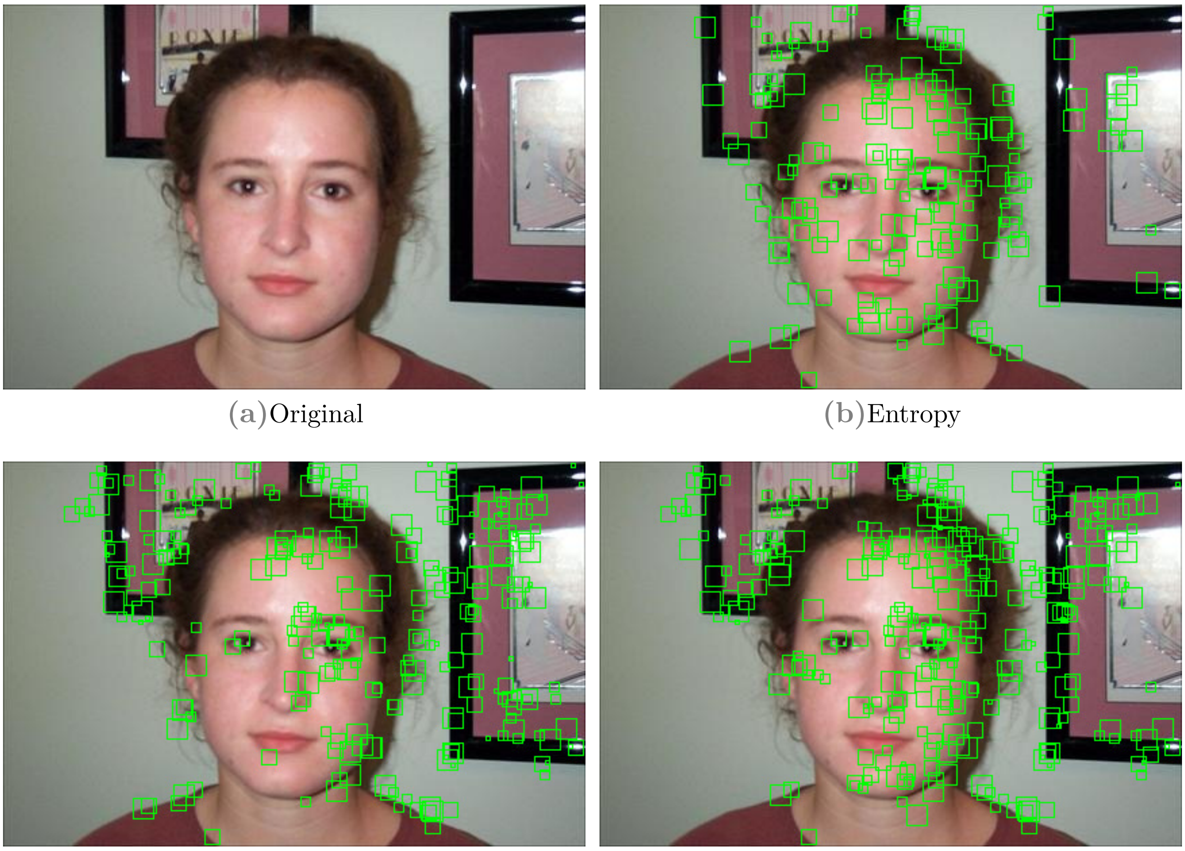
(a)Input (b)Standard (c)Enhanced

Figure 5.14 The figures at (b) show the unadaptive feed-forward template (blue spots) sampling of the standard HMAX model. Our approach (column c) adds LGN feedback with entropy-sensitive selection according to the template's information gain. It shows that templates sampled in areas with low information (like the surface of a street or a wall) are rejected. That way areas are selected which are easier to distinguish, which helps in the classification process. The pictures were chosen from the Caltech-101 database.

The intensity difference and the standard deviation approach were both equally fast but about 1.5× faster than the entropy approach, with similar results (see figure 5.15). In order to evaluate our entropy approximations, we ran the different approaches on a test set of random images. We randomly sampled templates and rejected those below a certain threshold, which was chosen for each approach individually. We then counted the remaining not rejected number of patches, calculated their average entropy and overall entropy. The results are displayed in table 5.3 and show that our approximation approaches can give a good estimate for entropy.

Table 5.3 Results for our different entropy approximations averaged over randomly chosen images. We sampled 800 templates and rejected those below a certain threshold for each approach. The number of not rejected templates is quite similar for each approach, as well as the overall sum of entropy of those templates and the average template entropy. It indicates, that our approximations can be used for estimating the information in a template. See figure 5.15 for a visual comparison.

	Min Max	Std Dev	Entropy
Not rejected Templates	239	230	234
Entropy Sum	905	870	922
Average Entropy per Template	3.758	3.752	3.89



(a)Original (b)Entropy

(c)Std Dev (d)Min Max

Figure 5.15 Approximations for calculating the entropy function. The green rectangles in figure (B) indicate areas with a high entropy. (C) shows the result with an approximation for entropy using standard deviation. (D) shows the result using just the difference of the maximum and minimum intensity. See table 5.3 for numerical comparison.

We tested our system against the Caltech-101 database. For each run, we randomly chose a training and testing image set and computed results with different numbers of positive training examples (1, 3, 15, 30 and 40) and 50 negative training examples. Our approach outperforms the original system in regard to the classification accuracy (e.g. for the airplanes dataset 92% compared to 86%; faces: 96% to 90%, cars 96% to 94%) or is at least of equal result (see figure 5.16).

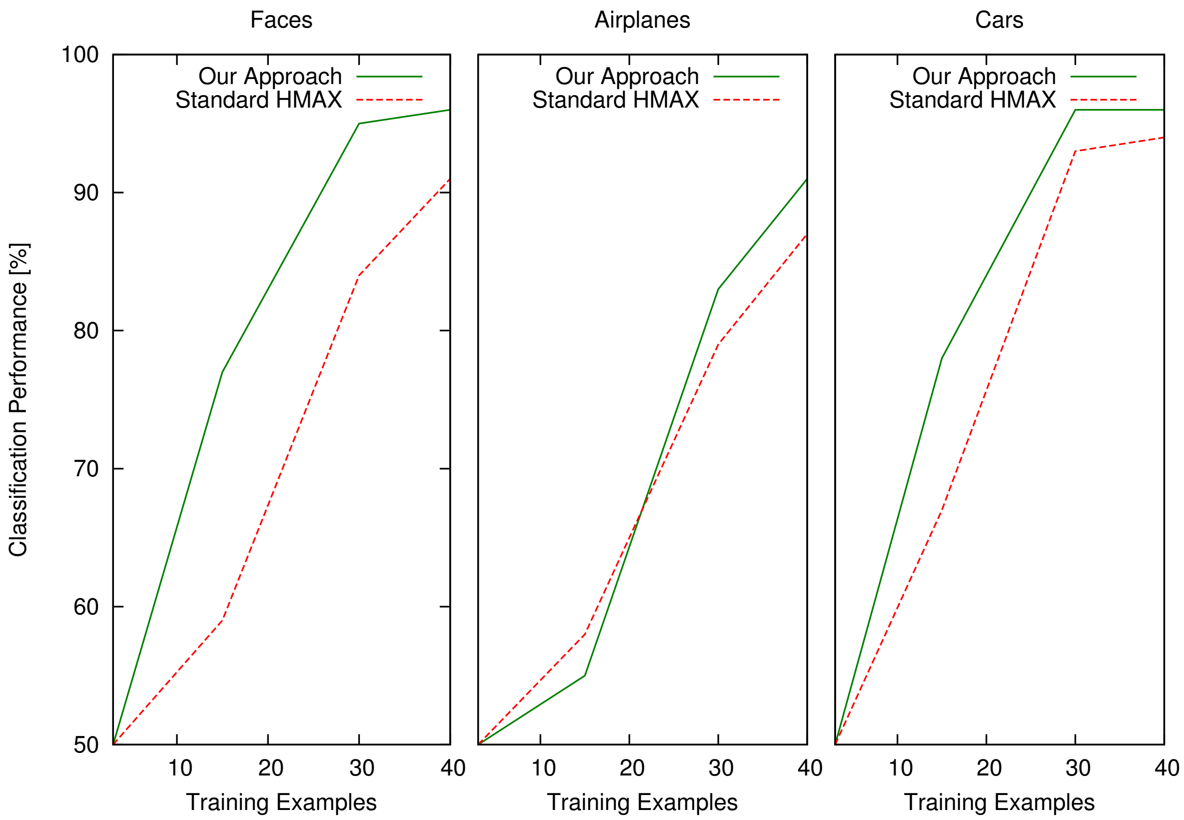


Figure 5.16 Comparison of classification results for faces, airplanes and cars of the Caltech image database between the standard HMAX and our entropy-enhanced model.

The complexity of calculating a template entropy is $O(c \times r)$ with c being the columns and r the number of rows of the template. The complexity for calculating n samples with t templates in the dictionary is $O(n \times t)$ and on a pixel basis $O(n(c \times r) \times t(c \times r))$. We can assume, that $t > (c \times r)$, which results in a lower overall complexity when applying our entropy approach. This is because we avoid calculating $t(c \times r)$ operations with each rejected low-entropy template.

5.3 Temporal Reasoning

In this section we focus on the temporal aspect of object recognition, which has only received little attention so far.

It is known that the entire process of object recognition activates much more areas in the brain than just the visual cortex, which indicates that in order to achieve similar efficiency in technical applications, simple feature generation and classification alone won't solve the problem in the long run. We need to regard object recognition not as a distinct but as a cognitive process. According to the principle of temporal contiguity a cognitive association is made between objects seen in rapid succession [Li and DiCarlo, 2010].

The HMAX model was developed as a proof-of-concept for reproducing the human performance in rapid scene classification - a test used in psychology and neuroscience to describe how fast and how well a human subject reacts to the task of distinguishing between categories in natural images shown for a very short period, where there's no time for eye movement of shifts of attention [Peelen et al., 2009; Li et al., 2002].

Classical object recognition systems in technical applications also disregard a temporal influence on the classification process. But in real world scenarios and especially in robotics, it is essential to model uncertainty and make use of the robot's abilities to act on it. With a humanoid robot and its active vision system and manipulators we have the tools to model the uncertainty by including the temporal aspect. Accounting for time could push current models from static single image recognition to a higher level of object consciousness

Rapid scene classification as applied in studies is usually not the normal way how humans perceive their environment [Deubel and Schneider, 1996]. To identify objects we usually move our eyes to different salient areas to gain some kind of certainty about our belief what the object might be [Farah, 1992]. Depending on the visibility of the scene, this procedure might vary in time until some certainty is gained [Goldstone, 1998]. Our system reproduces this behavior by applying a biologically-inspired object recognition model to a time-aware architecture.

Our approach uses a support vector machine classifier which supports probabilistic outputs of the membership likeliness for each trained class [Wu et al., 2004; Chang and

Lin, 2011]. The evaluation shows (figure 5.17), that these probability estimates can be used as a certainty measure of the right classification: If a feature vector is more likely to represent an object of e.g. class 1, the probability for that class is higher than for the other classes. We tested how representative and valuable those results are for being applied to our system. Figure 5.17 shows the probabilities' frequency distribution of a two-class classification benchmark over multiple runs. The green bars represent the true positive, the red bars the false negative test results. It shows, that the probabilities give a good estimate of how likely the class assignment is. The false negative votes had an average probability of about 63% whereas the true positive probability was about 87% (keep in mind: these probabilities are not the classification average, but the average of the probability responses of the true positive/false negative tests). We ran multiple test to verify the correctness of our probability over time approach.

Figure 5.17 Probabilities' frequency distribution of a two-class classification benchmark.

In terms of a time series of probabilities this allows us to make further assumptions:

1. If we see a high probability, we can assume less risk of false classification.

2. If we see a low probability, we're less certain, because we face a higher risk of a false negative test.

Therefore we assume, that the higher the response to a certain object and the more often this signal appears, the more likely it is that the response really represents the object we see. To model this behavior, we make use of psychophysics – a method used for example in digital signal processing and cognitive neuroscience to describe the likeliness of a perceptual system's response to a frequent stimulus. This approach is similar to concept of summation in neurophysiology. Summation is the method of signal transduction between neurons, which determines if an action potential will be triggered and the neuron fires. There exist two types of summation - spatial and temporal. Spatial summation refers to the achieving of the action potential by summing the input of multiple presynaptic neurons firing at roughly the same time. Temporal summation on the other hand refers to achieving the action potential by summing the potentials of a single presynaptic neuron firing frequently over a short period of time - in our case the frequent probability estimates of the different classes.

We apply probability summation over time [Watson, 1979] - a method used in signal detection theory - which models the probability P that a signal is detected accounting for all P_i, with P_i being the probability that a temporal stimuli threshold is exceeded at time i.

$$P = 1 - \prod_i (1 - P_i) \tag{5.15}$$

Equation (5.15) is the probability for one channel. We are interested in n channels, or in our case n classes, which compete to reach the threshold. Therefore we apply a maximum function over the set of classes k:

$$P = \max_k (1 - \prod_i (1 - P_{i,k})) \tag{5.16}$$

In our case we chose the threshold to be at least $\frac{1}{n} * 100\%$ - with n being the number of possible classes - to have a probability above 0% of getting detected, because if the probability is below the threshold, there exists at least one signal which has a higher value. We fit this constraint to an exponential distribution to model the assumption that higher probabilities are more likely to be correct than lower probabilities. This was empirically evaluated earlier in figure 5.17 and the true positive curve shows exponential character as well. We map the probability values from $[\frac{1}{n};1]$ to be in a range between $[a;b]$ in exponential space. Therefore we calculate

$$f(x,n) = \exp(g(x,n)) \tag{5.17}$$

with the mapping function

$$g(x,n) = mx + t \tag{5.18}$$

With our constraints

$$g(\tfrac{1}{n},n) = a; \quad g(1,n) = b; \tag{5.19}$$

with a being the lower bound and b the upper bound of the mapping, we get

$$m = \frac{b-a}{1-\frac{1}{n}}; \quad t = b - m; \tag{5.20}$$

We now need to normalize equation 5.17 to a range from $[\frac{1}{n};1]$ by applying

$$f(x,n) = \frac{\exp(g(x,n)) - exp(a)}{exp(m*(b-m)) - exp(a)} \tag{5.21}$$

Because Equation (5.21) is a continuous probability distribution defined for $[1/n;1]$, we set $P_{i,k,n} = f(R_{i,k},n)$ from Equation (5.15) with $R_{i,k}$ being the probability response for class k at time i from the classifier:

$$P = \max_k(1 - \prod_i(1 - f(R_{i,k}))) \tag{5.22}$$

Figure 5.18 visualizes the resulting graphs of equation 5.22 for $n = 2,3,4,5$. Figure 5.19 visualizes the probability estimates collected in our experiments and our estimated probability over time function for a two-class problem.

5.3.1 Accounting for Non-Static Scenes

Static object recognition assumes a constant discrete image. In our temporal approach we can't make that assumption, because in real-world scenarios the environment can

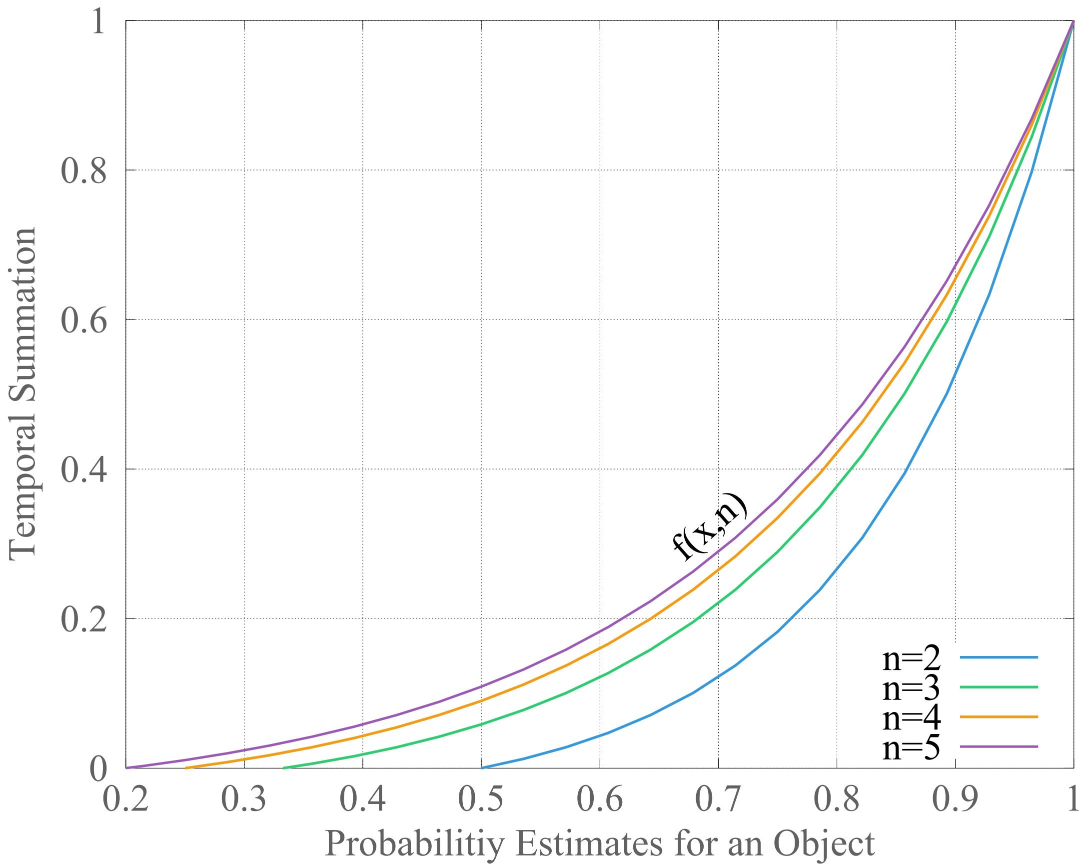


Figure 5.18 Stimuli functions from Equation 5.21 for different number of classes n.

change due to active influence like changing the point of view, or passive influence like an moving object. This makes it necessary to model the environment's uncertainty.

Objects in the visual field normally don't suddenly disappear or change its structure. According to the principle of temporal contiguity an association is made between objects seen in rapid succession [Li and DiCarlo, 2010]. Any difference would be interpreted as a displacement or masking of the object. Without the consideration of external motion or ego-motion, it would not be possible to build a believe system over an object in the visual field.

Therefore, we model motion as a trigger for resetting the classification believe certainty. If there is unsuspected motion from an external force e.g. something moves in the field of view or the object is taken away or replaced, we reset the believe probability in Equation (5.22) back to 0%. We model the motion detection as a stream separate from

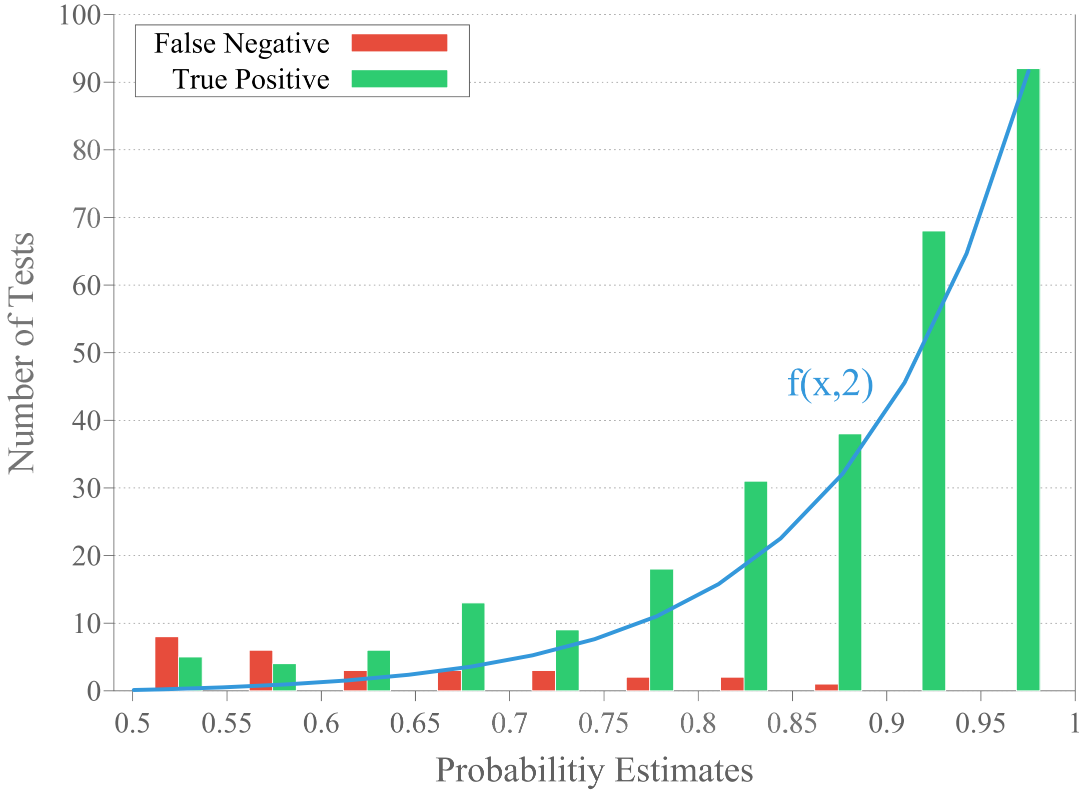


Figure 5.19 Comparison between estimated probability over time function $f(x, 2)$ (see equation 5.21) and collected data (shown in figure 5.17).

object recognition, which responses also end in the decision node. We detect the motion by reacting to a certain threshold to account for noise in the image data.

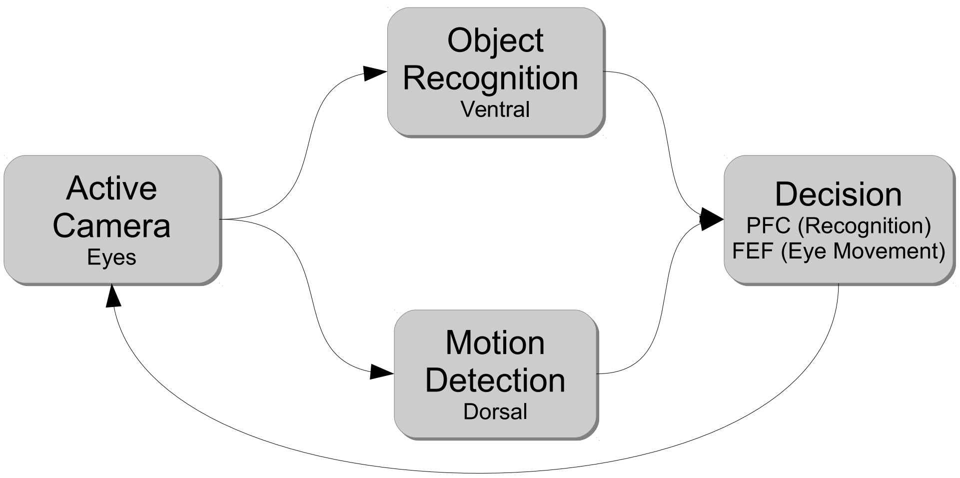


Figure 5.20 Architecture for Temporal Reasoning.

The images are processed in two parallel nodes, which handle object recognition and motion sensitivity - similar to the ventral and dorsal stream in the visual cortex (see figure 5.20). The responses of both stream are integrated into the decision making node - which is slightly based on the functionality of the prefrontal cortex (PFC) and the frontal eye fields (FEF) - an area located in the PFC, which is responsible for guiding eye movement and saccades. The decision node integrates the classification response probabilities over time and the external motion in the visual field to calculates a certainty measure over the present object. If after several trails the robot is still uncertain about the object, it could move its eyes or torso to a different position to have a better view point, or move or turn the object itself.

5.4 Processing Speed

We compared the computation speed for convolution with different filter sizes on CPU and GPU for the separable filter and the non-separable filter in figure 5.21 for a image size of 320×240. Using our separable filter approach we achieve a constant processing time on GPU of under 1 ms for $j = 3$ on all kernel sizes. The average computation time of the S1 layer using our ModHMAX approach with 4 Gabor filters takes under 8 ms on GPU compared to about 256 ms for 64 filters on CPU with the standard system (see table 5.4). This is a speed up of about 64.

Compared to our CPU implementation of the standard HMAX model with nonseparable Gabor filters, our system speeds up the computation using GPUs and separable orientation-free Gabor filters by a factor of $\approx \mathbf{16.8}$ (see table 5.4).

Table 5.4 Processing speed of S1 layer in HMAX vs our system (averaged over 100 cycles; CPU: i7, GPU: Geforce 670 GTX).

	HMAX		ModHMAX	
	CPU	GPU	CPU	GPU
Non-separable filter	252 ms	98 ms	32 ms	12 ms
Separable filter	177 ms	60 ms	22 ms	8 ms

In table 5.5 we show the computation speed for the next layer C1. Again we compared the speed of the original HMAX system against ours.

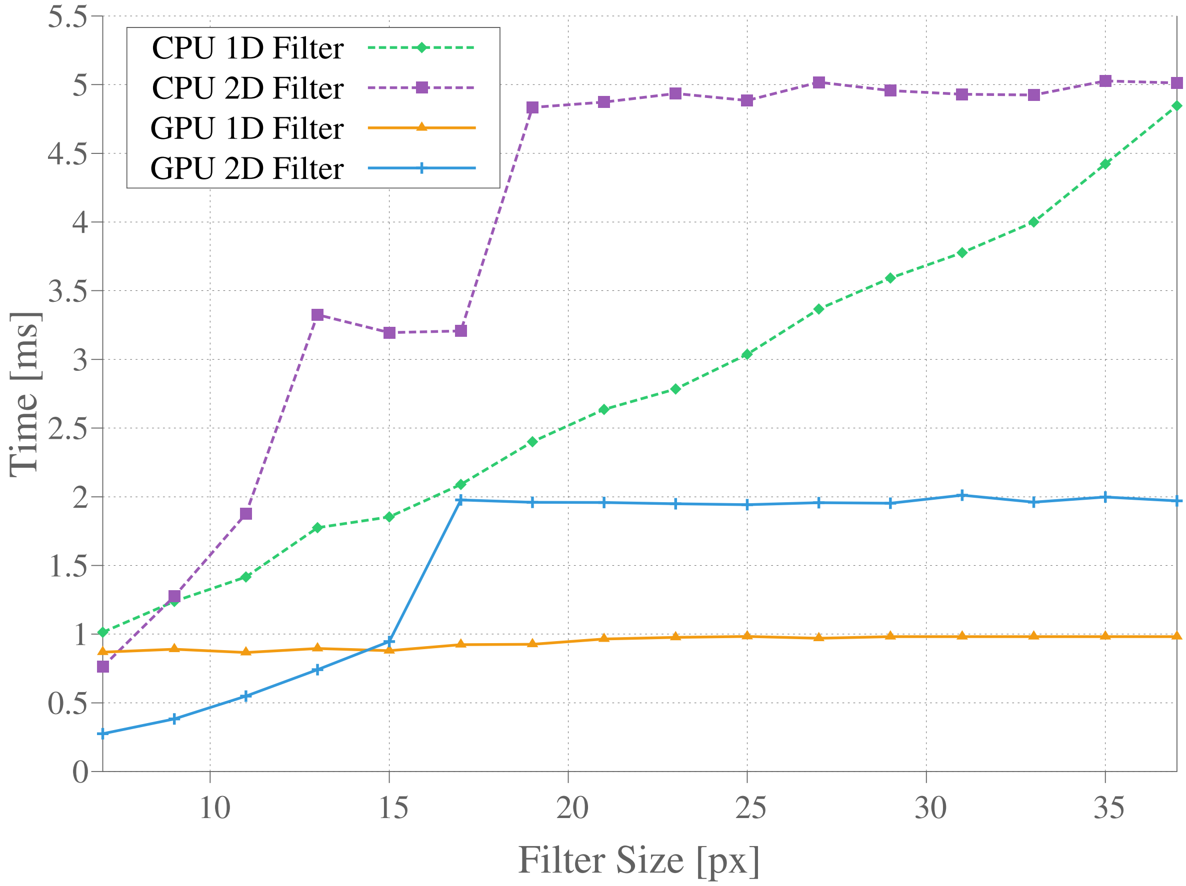


Figure 5.21 Speed comparison between Image Filtering with non-separable and separable kernel using CPU and GPU for different kernel sizes.

Table 5.5 Processing speed C1

	HMAX		ModHMAX	
	CPU	GPU	CPU	GPU
MAX Operation	140 ms	37 ms	17.5 ms	4.65 ms

Table 5.5 shows processing speed for a dictionary of size 2000 with a sampling rate of 200 patches per patch size per C1 layer. Our system speeds up the overall processing for the S2 Layer by a factor of ≈ 8.6.

5.5 Summary

In this chapter we presented our ModHMAX object recognition system - a modified and enhanced version of HMAX with a focus on time-crucial applications for real-world scenarios. We were able to speed up the first two layers using singular value decomposition and GPU programming. We evaluated the different combinations of Gabor filters and Max Pooling Size. Additionally, we investigated the different effects of dictionary and sampling size to the classification accuracy. The building of the dictionary was modified so that object-specific dictionaries can be built for localization purposes. We also introduced an information-theoretic measure into the system using an entropy metric, which is able to improve classification accuracy by up to 6%. Finally we presented a temporal reasoning approach which enables the system to built a believe system over time, to mimic a more realistic approach to biologically-inspired object recognition.

In the next chapter we present an approach towards the integration of the previously presented fields, namely visual attention, object-based attention and object recognition. We suggest a software architecture and apply it on the humanoid robot iCub.

Chapter 6

A SYSTEM ARCHITECTURE FOR VISUAL ATTENTION, OBJECT SEGREGATION AND OBJECT RECOGNITION

This chapter proposes an architecture for the integration of visual attention, object-based attention and object recognition for active camera systems. We describe how the single modules are integrated into a software framework and how the communication and information processing is handled between the modules. We successfully enable the humanoid robot iCub to adjust the gaze to the most salient point using out visual attention system based on sampled template collation. The fixated object is then fed to our object-based attention system for object segregation. The segmented object is then classified using our ModHMAX approach.

First we describe the general idea of the software architecture and how the single modules - visual attention, object-based attention, object recognition and temporal reasoning - are connected and what information they exchange. In the second section we explain the realization of the architecture in the software frameworks ROS and Yarp on the humanoid robot iCub.

6.1 Biologically-inspired foundation

The architecture and functionality of our system was inspired by the highly parallel anatomy of the human visual system and the way information is processed between the neurons [Nassi and Callaway, 2009]. Following the biological model, the computational architecture's focus lies on multiple strategies: The hierarchical processing [Felleman, D.J. and Van Essen, 1991], the parallel processing [Nassi and Callaway, 2009], the modularity and the plasticity [Kourtzi et al., 2006; Lomber et al., 2010].

A large body of works have been proposed recently with a strong relation to neuromorphic algorithms and their implementations [Indiveri et al., 2011; Kourtzi et al., 2006; Rachmuth et al., 2011]. However, most of this research focuses on building specific models e.g. Itti and Koch's computational model of visual attention [Itti et al., 1998], which does not necessarily make them applicable in real-world scenarios. Other projects (e.g. blue brain project [Markram, 2006; Maass et al., 2002]) make use of large scale computing to simulate and visualize neuronal models with a focus on the very detail of even ion channel distributions, but do not consider its functional utilization.

In our work, we propose a functional architecture, which can spread the computational processing streams over a cluster of PCs with multiple CPUs[1] and GPUs[2] using a construction of software nodes that are responsible for the functionality. The software architecture is based on the well-established robot operating framework (*www.ros.org*) which handles the communication and synchronization between the single processes.

We created our computational architecture corresponding to the functionality and structure of the human brain by embracing following key-features:

[1]Central Processing Unit
[2]Graphics Processing Unit; highly parallel structure

Plasticity

The brain is able to change neural pathways, e.g. by experience or brain damage [Kolb and Whishaw, 1998; Voytek et al., 2010], and also between different brain regions [King et al., 1988; Finney et al., 2001].

Due to the general approach, the architecture is able to dynamically change its behavior and it's connections among the different nodes. It is possible to cope with possible loss of a processing unit by redirecting the processing stream and spreading the same functionality over multiple PCs gaining redundancy of functionality and data. The architecture is not dependent on a certain position, the computation can be spread over any number of PCs at any possible place as long as they are connected via Ethernet.

Modularity

The anatomical organization and functional organization of the brain is modular. Different areas in the brain cope with different tasks, e.g. auditory or visual [Chen et al., 2008]. Each area is defined by dense internal connectivity and relatively sparse external connectivity to other modules [Unit, 2009].

Our architecture is modular and expandable both in hard- and software. The software nodes are not bound to run on a specific computer and can be organized according to their computational complexity. The hardware can be extended within a single PC, e.g. with a better GPU; or with another PC. The functionality can be broken down to into several processes to spread the computational payload.

Connectivity

However far from fully understood, different regions in the brain are functionally connected to enable sensory integration and perform motor and cognitive tasks [Horwitz et al., 2003; Rubinov and Sporns, 2010]. This is realized by functional integration and dynamic interaction within the different areas and neurons [Sporns et al., 2000; Breakspear, 2004].

The nodes can be connected dynamically during runtime. The data is streamed via network using TCP/IP[1] or for larger data UDP[2] for lower latency. For several technical reasons we primarily utilize the UDP protocol:

1. as less data needs to be transferred, due to a lower overhead compared to TCP;
2. If data is lost we do not require it to be resend (which is the case for TCP), as older data are of no interest in our architecture, but the current one, because we emphasize the importance of the immediate sensory awareness of the surrounding world;
3. If the arriving information is outdated, the system won't use it for further processing. The discrepancy of asynchronous data could lead to false results.
4. To achieve this level of reliability TCP would need to send data back to the sender which negatively affects the bandwidth and can increase the network's latency.

Concurrent and Parallel Processing

Information processing happens in highly parallel fashion in the brain and is essential for coping with the large input of the sensory information and the mammal's ability to quickly react to sudden threats. Multiple processing stream, each more or less responsible for a particular type of low-level sensory cue, process the sensory information at the same time [DeYoe and Van Essen, 1988; Rauschecker, 1998; Ballard, 1986].

In our system the data is processed in parallel using multiple CPUs, GPUs, and PCs. The degree of synchronization between the single processing streams can be chosen manually or even dynamically. The synchronization policy can be chosen to be exact, within a time frame, or approximate. Thus, we can ensure that the processed data is from the same time period.

Hierarchical Processing

Besides the parallel processing in the brain, the information is handled in a hierarchical way, because different regions depend on the functional output of other

[1]Transport Control Protocol / Internet Protocol; uses retransmission in case of message loss
[2]User Datagram Protocol; unidirectional transmission

areas, e.g. in the visual cortex, where information is passed on from V1 to V2 and V4 [Felleman, D.J. and Van Essen, 1991; Bodegård et al., 2001].

A hierarchical processing is realized using the nodes and connections between them on multiple layers with different synchronization strategies. This results in a time-window-synchronous processing between the layers, whereas the processing within the layers is highly parallel.

Figure 6.1 Experimental set-up. A computer cluster and monitors showing a running ModHMAX system.

Figure 6.1 shows an early experimental set-up of our architecture running ModHMAX. We created a cluster consisting of four PCs; two with an Intel i5 quad core CPUs and one PC with two Xeon quad core CPUs. One PC with an i7 is also equipped with a

Nvidia GeForce GTX 580 with 560 CUDA Cores with a processor clock of 1.5 GHz. We connected the PCs to six monitors to visualize the processed data.

6.2 Active Camera Systems

In an active camera system, the viewpoint of the cameras can actively be manipulated, for example by using a pan and tilt unit which provides two DOF.

From a technical perspective, active vision can be used to improve perception in a couple of cases [Aloimonos et al., 1988; Rivlin and Rotstein, 2000]:

- Tracking of objects - simply by the ability to follow the trajectory, but also by a reduction of motion blur evoked by exposure and the rapid movement of the object.
- Occluded Objects - the point of view can be changed to avoid the occlusion.
- Limited field of view
- Limited resolution of the camera
- Enhancing depth perception
- Reduction of sensory information - using one active camera instead of multiple static ones.

From a psychological perspective, active vision systems could also be beneficial during human-robot interaction when integrated in a robot head. A robot that looks and behaves more human-like is more likely to be accepted by a person. Especially eye contact is a social form of non-verbal communication and has a strong influence on our behavior. They provide some sort of social and emotional information of the other person and play a major role in facial expressions, which can indicate someone's emotional state and intentions. Talking to a person with sunglasses is for example a situation, where no real eye contact can be made that produces a lack of non-verbal communication, which can make people feel uncomfortable. We can help in preventing an alienation of robots in close human-robot environments by using active camera systems as robot eyes.

6.3 General Software Architecture

The active camera system provides images to all modules - visual attention (VA), object-based attention (OBA) and object recognition (OR). The VA module uses the image to calculate the most salient point in pixel position. The coordinates are then fed to the active camera module to control the gaze position and to the OBA module which uses them to calculate the OBA map using the image from the camera. The resulting OBA map is sent to the OR module, which internally creates the feature maps and samples only from the area suggested by the OBA map to create the classification probabilities. These probabilities are then used by the temporal reasoning model to infer the object class over time. If the eyes move or if there is any larger movement in the image, which might indicate that the fixated object is not longer in the image, the internal believe system of the temporal reasoning module is reset in order to avoid false classification. Figure 6.2 gives a simplified overview of the different modules and their connections.

6.4 The Humanoid Robot iCub

The iCub is a humanoid robot that was built to study cognition and designed to resemble the looks of a 3.5 year old child [Metta et al., 2008]. It was created by the RobotCub project, a 5 years long project funded by the European Commission through Unit E5 "Cognitive Systems, Interaction & Robotics". The iCub has a 6 DoF head (3 DoF for the neck, 3 DoF for the cameras (eyes)). The active eye cameras make the iCub especially suitable for research related to human vision.

To control the iCub gaze and saccades we use iKinGazeCtrl, a iCub module based in iKin - a library for forward and inverse kinematics [Pattacini, 2011]. iKinGazeCtrl provides functions for saccades steering the neck and the eyes independently. We supply it with the pixel coordinates of the most salient point.

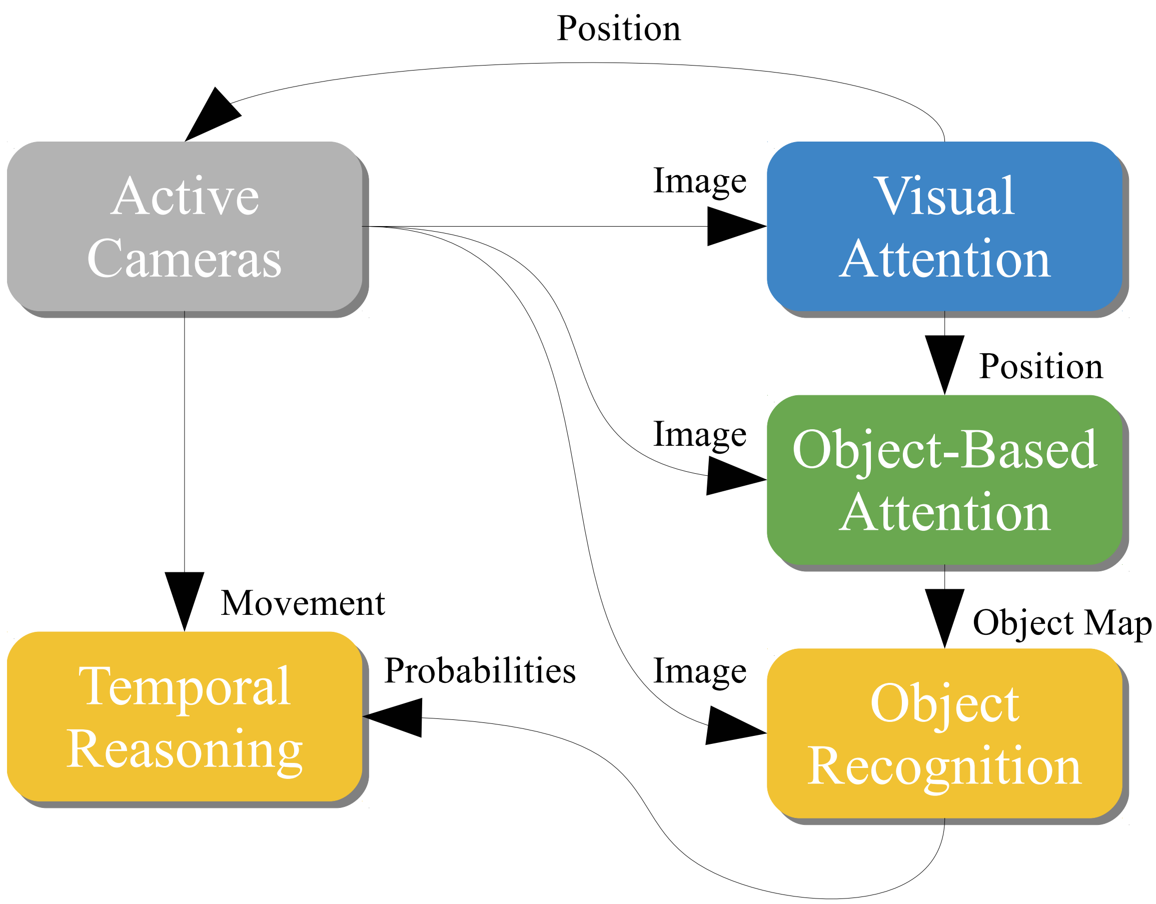


Figure 6.2 Simplified overview of the system architecture.

Figure 6.3 The humanoid robot iCub. [Metta et al., 2008]

6.5 Software Integration

We integrate our modules in ROS - the robot operating system. It provides libraries and tools for message passing, message synchronization and parallel processing over one or multiple PCs using nodes. A ROS node is a process that runs on a PC and can communicate to other nodes using TCP/IP or UDP. The connection between nodes is established using a mater - the ROS core - which acts as a nameserver. The messages are published by using an identifier called topics, which is made known to the ROS core so that other nodes can subscribe to it.

The framework applied with the iCub is Yarp - Yet Another Robot Platform [Metta et al., 2006]. Like ROS it provides tools for message passing between processes and devices. Instead of topic, Yarp messages are identified by ports.

6.5.1 The Main Modules

Here we describe the single modules of our architecture (see figure 6.4):

- The iCub modules are controlled using Yarp. The eye camera provides its images on a Yarp port called `/icub/cam`. We integrate the functionality to access the Yarp network and the ports in a ROS node `/yarpToRosImage` and subscribe to the Yarp port in order to obtain the camera image.
- The visual attention node (blue) subscribes to the image message published by the `/yarpToRosImage` node. It then calculates the saliency map using our sampled template collation approach and publishes the most salient point in pixel coordinates.
- The object-based attention map (green) subscribes to the pixel coordinates of the most salient point from the visual attention node, calculates the object-based attention map and publishes it.
- The object recognition and temporal reasoning modules are arranged in three different nodes (yellow) for reasons of efficiency and modularity.
  - The `/S1_C1_cuda` node subscribes to the camera image and calculate the first and second layer of our object recognition system. The convolution with the

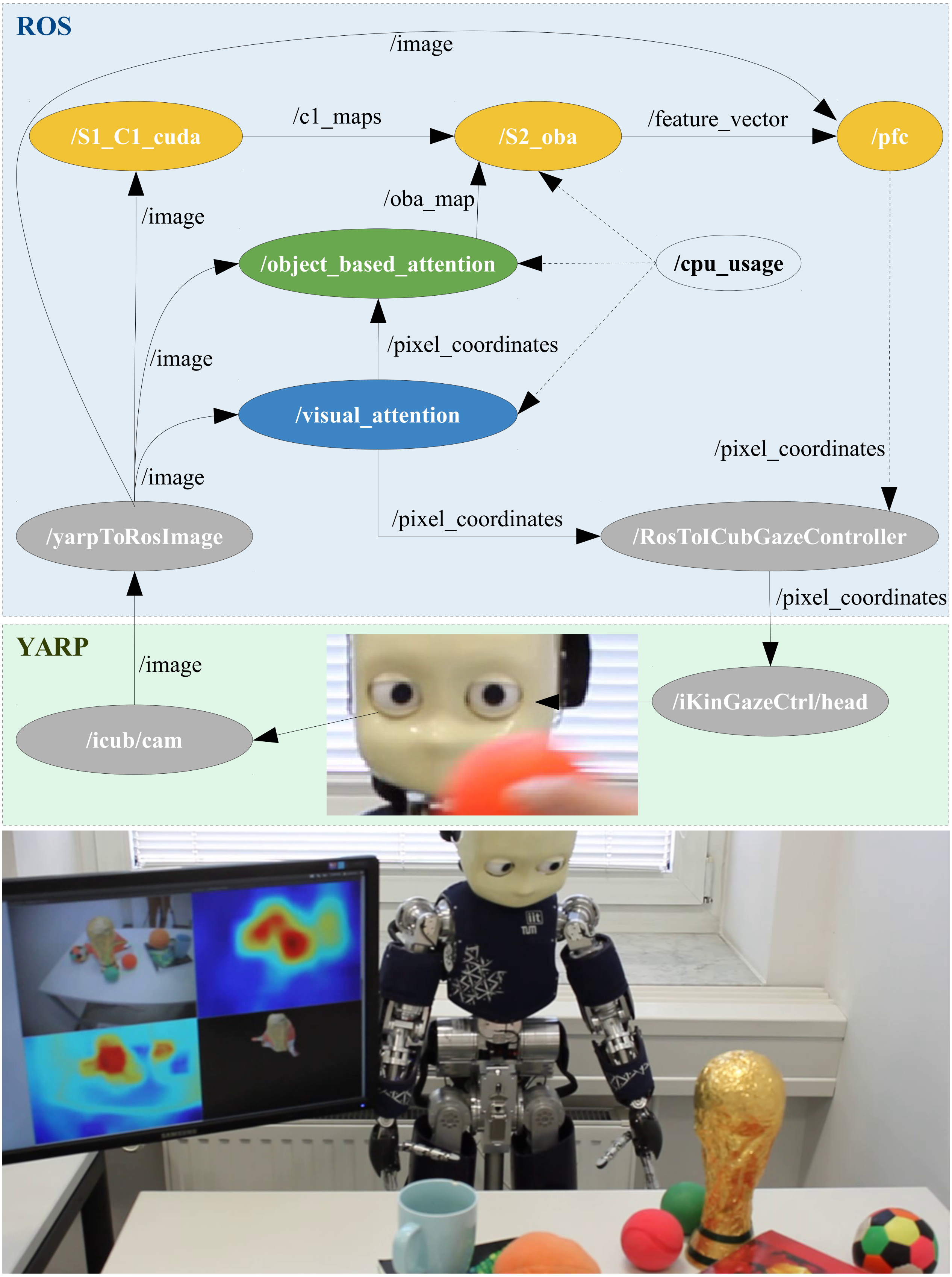


Figure 6.4 Our architecture using ROS and YARP running on the humanoid robot iCub.

Gabor filter and the Max Pooling operation are calculated on GPU using CUDA. This way the node can be run separately on a PC with a suitable graphics card. After the calculations the node publishes the resulting C1 maps.

- The `/S2_oba` node integrates the S2 and C2 layer of the object recognition system. It subscribes to the C1 maps and to the object-based attention map. It samples templates from the C1 maps only from those areas which are suggested by the object-based attention map. Then the feature vector is generated by calculating the maximum response of the templates to a set of initially sampled templates - the dictionary. This feature vector is then published.
- The prefrontal cortex node `/pfc` integrates the classification, the temporal reasoning and the motion detection. It subscribes to the feature vector supplied by the S2 node to classify the object, the classification output is then used for our classification over time approach. It also subscribes to the camera image to detect if there's any movement to reset the internal believe system. The pfc node can control the eye movement, which is necessary to keep the fixation on the object as long as it is not properly classified with certainty, or to slightly change the fixation to get a better result. Otherwise the visual attention module would change the fixation to the most salient point.

- The `/RosToICubGazeController` node subscribes to pixel coordinates of the visual attention node and the pfc node. It also integrates Yarp functionality to send the pixel coordinates to the `/iKinGazeCtrl/head` port in the iCub's Yarp network.

6.5.2 CPU Usage and Synchronization

In time-crucial scenarios where soft real-time is desired, like in robotics, it can be necessary to look at the CPU consumption of the system. If the processor operates under full load, it could happen that the system becomes non-reactive and we risk that certain processes freeze and stop working properly. This can become a safety hazard especially in environments with human-robot interaction.

We therefore introduced a node into the system that measures the CPU usage on the PC. The `/cpu_usage` node publishes the percentage of CPU load in the system. The visual attention and object-based attention node subscribe to it and can adjust the sampling rate accordingly. We can set a lower bound to maintain an amount of stability of the results, as evaluated in the previous chapters. The `/S2_oba` node also can subscribe to the cpu load node and adjust the sampling rate of the templates sampled from the C1 maps. A lower bound is however in this case harder to find as it depends on the number of objects trained, the size of the dictionary and the size of the segmented object.

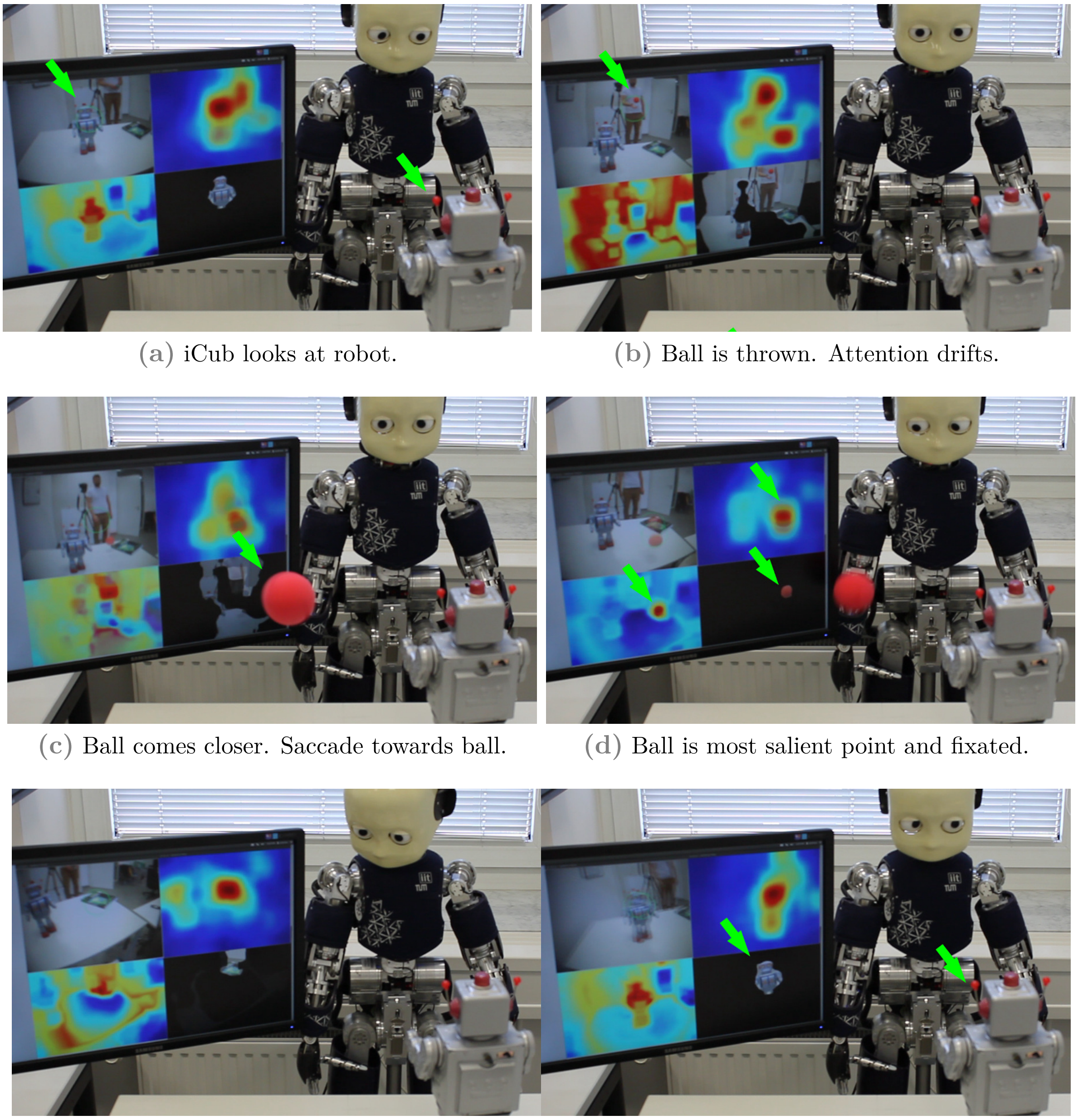
In real-world scenarios it is important to handle message synchronization to avoid non-consistent information about the environment. If the processed data is outdated, the produced information is not only obsolete and redundant but might also cause action that endanger the environment. ROS provides some tools to tackle this problem:

- The Time Sequencer can be used to ensure that messages are processed in temporal order according to the header's timestamp, which helps to process only the most up-to-date data.
- The Synchronizer is used if a node has subscribed to multiple messages which share a common callback function. The object-based attention node subscribes for example to the camera image and the pixel coordinates of the most salient point. If the pixel coordinate message is delayed for too long, the callback function would produce a map which is based on outdated data. In our architecture we use the ApproximateTime policy which matches the messages according to their time stamp, allowing some time difference.

6.5.3 Processing

Figure 6.6 gives an overview of the overall processing. First an image is captured from the iCubs active camera. This image is sent to three nodes: Visual Attention, Object-based Attention and Object Recognition (in our case `/S1_C1_cuda`, the node for the first two layers of our recognition system). Additionaly the image is sent to the `/pfc` node which handles the temporal reasoning.

The visual attention node now calculates the most salient point in the image and sends the image coordinates to `/object_based_attention` and to `/RosToICubGazeController`.



(a) iCub looks at robot.

(b) Ball is thrown. Attention drifts.

(c) Ball comes closer. Saccade towards ball.

(d) Ball is most salient point and fixated.

(e) Ball out of sight

(f) iCub looks back at robot

Figure 6.5 A test case to evaluate if our system is fast and efficient enough to detect a fast moving object. At first the robot is the most salient area to the iCub (a). Then a red ball is thrown, which slightly draws away the attention of the iCub towards the ball (b-c). At (d) the ball is fixated and correctly segmented with our object-based attention approach. In (e-f) the ball is out of sight and the iCub looks back at the robot. Note that there is no motion detection involved, just our Sampled Template Collation approach for Visual Attention and Object-based Attention.

The `/RosToICubGazeController` node converts the pixel coordinates to a yarp message and sends it to `/iKinGazeCtrl/head` which adjusts the iCub to center the eyes at the given pixel coordinates.

The `/object_based_attention` node calculates the object-based attention map using a current input image and the pixel coordinates from the visual attention module to sample a seed template from that position. The map is then sent to the `/S2_oba` node from the object recognition system.

The `/S2_oba` node gets the calculated C1 maps from the `/S1_C1_cuda` node and the object-based attention map and only samples templates in the C1 maps from the object area. It calculates the feature vector and sends it to the `/pfc` node.

The `/pfc` node uses the feature vector to classify the object and builds the internal believe system of the object class for temporal reasoning. If motion is detected using the input image, this believe system is reset.

Each node constantly processes new arriving data. The visual attention node calculates for each input image a new most salient point and tries to adapt to the frequency of the camera, which has about 60 Hz. We tuned the node to update the most salient point at about 20 Hz for the gaze controller, so that the active camera image is more stable. The object based attention map is updated at about the same frequency.

Figure 6.5 shows a test case with the iCub and a thrown ball to display how fast and efficient the visual attention and object-based attention system are running.

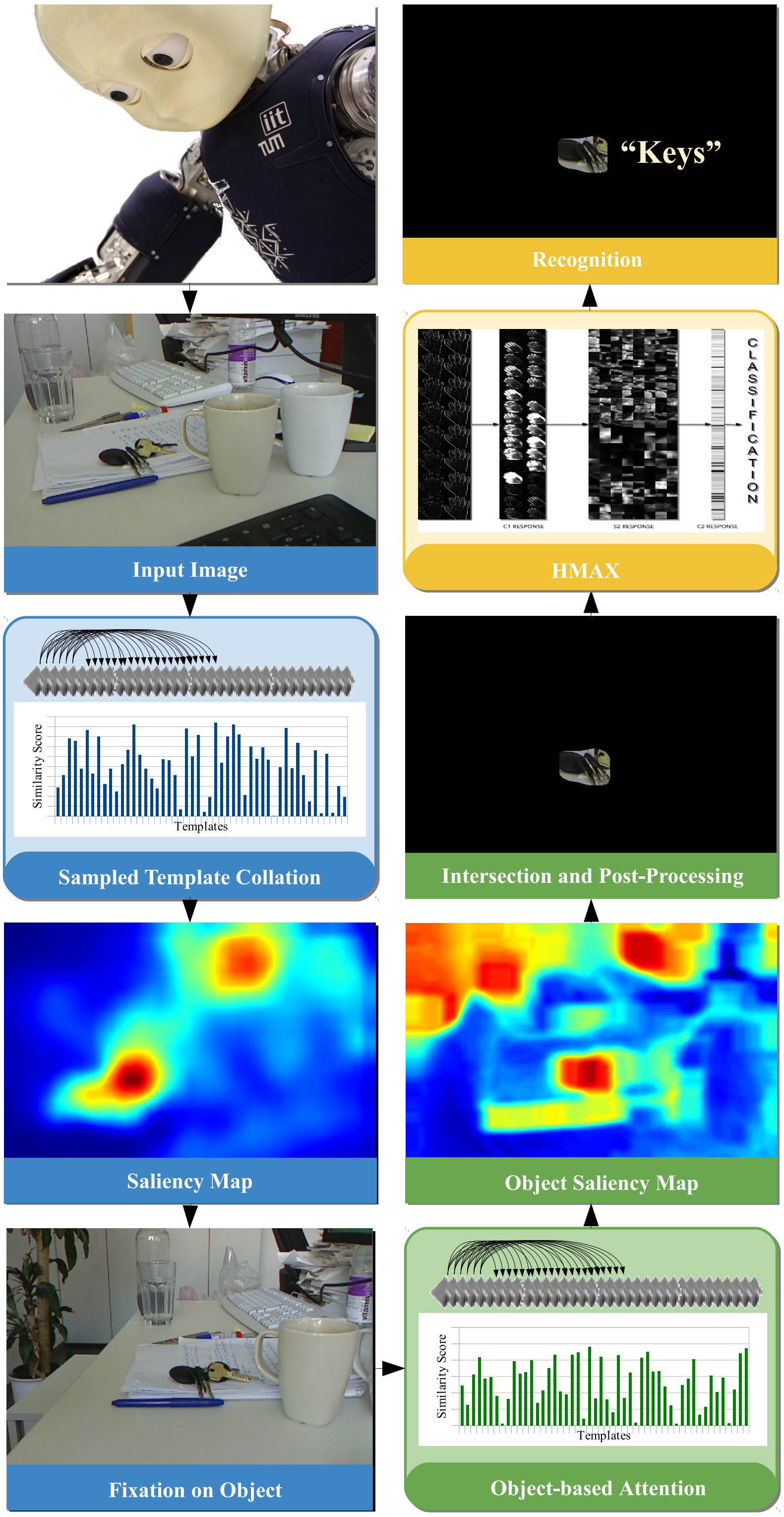


Figure 6.6 Processing Overview. First a saliency map of the acquired image is calculated using sampled templates collation (STC), then the most salient area is fixated with an active camera. The focused object area is then segmented using a STC-based approach to object-based attention. After eliminating areas that don't contain the object, the resulting map is used to subsample templates for object recognition.

6.6 Summary

In this chapter we proposed a software architecture for visual attention, object-based attention and object recognition for active camera systems. We implemented the architecture using ROS and Yarp and evaluated it using the humanoid robot iCub. The single modules run in parallel processes and communicate with synchronized messages. The architecture also supports dynamically adaption of the sampling rate in relation to the overall CPU usage, which helps in preventing a frozen and non-reactive system.

We showed that the processing is fast and efficient and was able to detect and segment even a thrown ball (see figure 6.5).

Chapter 7

CONCLUSION

In this chapter, we will first give a brief summary of the contents of this thesis. Then we will explain the contributions of this work. In the final section we will give an outlook of future work.

7.1 Summary

This thesis presented an approach towards the efficient integration of neuroscientific knowledge into a technical environment for improving vision models in time-crucial real-world scenarios in the context of humanoid robotics. We showed that by following the biological paradigm technical systems can be enhanced.

In chapter 3 we introduced a new method for generating visual attention maps and most salient point. The approach is based on our sampled template collation concept, which provides a fast and scalable way of calculation higher-level similarity relationships between regions of an image.

We extend this approach in chapter 4 to adapt the lesser known theory of object-based attention for technical applicability. We gave two application scenarios - 1.) Object-based attention for visual search and 2.) Object-based attention for object segmentation.

In chapter 5 we presented ModHMAX, a enhanced modification of the computational model HMAX for time-crucial applications like robotics. We introduced a new approach towards temporal reasoning in object recognition and presented an method to generate object-specific dictionaries for object localization.

In chapter 6 we presented a software architecture which integrates visual attention, object-based attention and object recognition for active camera systems. We tested our system on the humanoid robot iCub.

7.2 Contributions

The main contribution of this thesis is the analysis, the development and the evaluation of an efficient biologically-inspired vision system which supplies a humanoid robot with the ability to visually perceive and understand its environment in an efficient and scalable manner. We showed that by following the biological paradigm technical systems can be enhanced.

The proposed model integrates three essential parts of human vision: Visual attention, object-based attention and object recognition. Visual and object-based attention play a major role in how and what we perceive in our field of vision by selecting and reducing the available information. Both of which are essential for a fast and reactive vision system.

Six major contributions of this thesis helped to build this model:

- The design of a new visual attention system, which outperforms state-of-the-art system in terms of accuracy, speed and complexity realized by
- Our Sampled Template Collation method for efficiently evaluating different image regions, which is able to adapt to computational needs;
- A new object-based attention system, which enhances object recognition;
- Our object recognition model - an enhancement of a computational model called HMAX, which is an abstraction of the neural information processing in the visual cortex. The model was modified in terms of speed and performance in order to put it in a more technical context. We quantitatively and qualitatively show that by the integration of neuroscientific knowledge about neural information processing in the brain like lateral-inhibition and avoidance of entropic redundancy results in a higher classification accuracy and faster processing speed.
- The development of a temporal reasoning framework which enables the system to classify over time and account for uncertainties in non-static real-world scenarios.
- The system architecture for the efficient integration of visual attention, object-based attention and object recognition, which enabled the humanoid robot iCub to detect and segment even fast moving objects like a thrown ball.

7.3 Outlook

Although the presented work is complete in itself, as it integrates all parts into a working system, there is always potential for extensions and improvement. Here we want to suggest a couple of ideas for pursuing our approach:

- **Sampled Template Collation** could benefit from a more complex similarity calculation, like including the orientation of the templates.
- **Visual Attention** could be enhanced using a top-down approach in order to detect objects that our strongly biased in human visual processing, for example faces.
- **Object-based Attention** could be enhanced by using more than one seed template for the segregation process.
- **Object Recognition** based on HMAX has - like all other approaches its limitations: The more objects are learned, the harder it is to classify. It is our believe, that HMAX could benefit from more hierarchical layers and feature learning in the architecture itself. Which would be similar to recent deep learning approaches.
- **Using 3D information** would push the usability to more robotic related tasks and would enable an interaction with the environment. It also would be beneficial for visual attention and object-based attention.

Bibliography

Siddharth Advani, John Sustersic, Kevin Irick, and Vijaykrishnan Narayanan. A multi-resolution saliency framework to drive foveation. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2596–2600. IEEE, 2013.

John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4):333–356, 1988.

J.S. Anderson, I. Lampl, D.C. Gillespie, and D. Ferster. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, 290(5498): 1968–1972, 2000.

Alexander Andreopoulos, Stephan Hasler, Heiko Wersing, Herbert Janssen, John K Tsotsos, and Edgar Korner. Active 3d object localization using a humanoid robot. *Robotics, IEEE Transactions on*, 27(1):47–64, 2011.

D.H. Ballard. Cortical connections and parallel processing: Structure and function. *Behavioral and Brain Sciences*, 9(1):67–120, 1986.

Yoseph Bar-Cohen and Cynthia Breazeal. Biologically inspired intelligent robots. In *Smart Structures and Materials*, pages 14–20. International Society for Optics and Photonics, 2003.

Robert Bevec and Aleš Ude. The acquisition of visual representations for object recognition by autonomous pushing. In *21th International Workshop on Robotics*, page 156. Esa, 2012.

Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

A. Bodegård, S. Geyer, C. Grefkes, K. Zilles, and P.E. Roland. Hierarchical processing of tactile shape in the human brain. *Neuron*, 31(2):317–328, 2001.

MC Booth and E.T. Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8(6):510–523, 1998.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. 2013.

Michael Breakspear. âĂIJdynamicâĂİ connectivity in neural systems. *Neuroinformatics*, 2:205–224, 2004. ISSN 1539-2791.

Farran Briggs and W. Martin Usrey. Corticogeniculate feedback and visual processing in the primate. *The Journal of Physiology*, 589(1):33–40, 2011. ISSN 1469-7793.

Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2005.

Claus Bundesen, Thomas Habekost, and Søren Kyllingsbæk. A neural theory of visual attention: bridging cognition and neurophysiology. *Psychological review*, 112(2):291, 2005.

Marisa Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484 – 1525, 2011. ISSN 0042-6989. Vision Research 50th Anniversary Issue: Part 2.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

Shengyong Chen, Youfu Li, and Ngai Ming Kwok. Active vision in robotic systems: A survey of recent developments. *The International Journal of Robotics Research*, 30 (11):1343–1377, 2011.

Z.J. Chen, Y. He, P. Rosa-Neto, J. Germann, and A.C. Evans. Revealing modular architecture of human brain structural networks by using cortical thickness from mri. *Cerebral cortex*, 18(10):2374–2381, 2008.

Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *IEEE CVPR*, pages 409–416, 2011.

Carlo Ciliberto, Sean Ryan Fanello, Matteo Santoro, Lorenzo Natale, Giorgio Metta, and Lorenzo Rosasco. On the impact of learning hierarchical representations for visual recognition in robotics. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3759–3764. IEEE, 2013.

EH Cohen and F Tong. Neural Mechanisms of Object-Based Attention. *Cerebral cortex (New York, N.Y. : 1991)*, November 2013. ISSN 1460-2199.

Charles E. Connor, Howard E. Egeth, and Steven Yantis. Visual attention: Bottom-up versus top-down. *Current Biology*, 14(19):R850 – R852, 2004. ISSN 0960-9822.

R.L. De Valois, D.G. Albrecht, and L.G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, 22(5):545–559, 1982.

R Desimone and J Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18:193–222, January 1995. ISSN 0147-006X.

Robert Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1245–1255, 1998.

Heiner Deubel and Werner X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36(12):1827 – 1837, 1996. ISSN 0042-6989.

E.A. DeYoe and D.C. Van Essen. Concurrent processing streams in monkey visual cortex. *Trends in neurosciences*, 11(5):219–226, 1988.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–34, 2 2012. ISSN 1097-4199.

J.R. Duhamel, F. Bremmer, S. BenHamed, W. Graf, et al. Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389(6653):845–848, 1997.

Eilen Nordlie Eivind Norheim, John Wyller and Gaute T. Einevoll. Feedback and feed-forward contributions to temporal signal processing in the lateral geniculate nucleus. *Front. Neur. Conference Abstract: Neuroinformatics*, 2009.

Erkut Erdem and Aykut Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):1–20, 2013.

Martha J Farah. Is an object an object an object? cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science*, 1(5):164–169, 1992.

Felleman, D.J. and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47, 1991.

E.M. Finney, I. Fine, K.R. Dobkins, et al. Visual stimuli activate auditory cortex in the deaf. *Nature neuroscience*, 4:1171–1174, 2001.

Simone Frintrop. General object tracking with a component-based target descriptor. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 4531–4536. IEEE, 2010.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume*, volume 15, pages 315–323, 2011.

Christian Goerick, Heiko Wersing, Inna Mikhailova, and Mark Dunn. Peripersonal space and object recognition for humanoids. In *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pages 387–392. IEEE, 2005.

Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012.

Robert L Goldstone. Perceptual learning. *Annual review of psychology*, 49(1):585–612, 1998.

Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Image Processing, IEEE Transactions on*, 19(1):185–198, 2010.

Edmund T. Rolls Gustavo Deco. A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, 2004.

Tim Halverson and Anthony J Hornof. A computational model of âĂIJactive visionâĂİ for visual search in human–computer interaction. *Human–Computer Interaction*, 26 (4):285–314, 2012.

Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

B. Horwitz et al. The elusive concept of brain connectivity. *Neuroimage*, 19(2):466–470, 2003.

D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215, 1968.

Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210 – 1224, 2012. ISSN 0896-6273.

G. Indiveri, B. Linares-Barranco, T.J. Hamilton, A. Van Schaik, R. Etienne-Cummings, T. Delbruck, S.C. Liu, P. Dudek, P. Häfliger, S. Renaud, et al. Neuromorphic silicon neuron circuits. *Frontiers in neuroscience*, 5, 2011.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20 (11):1254–1259, 1998.

Laurent Itti. *Models of bottom-up and top-down visual attention.* PhD thesis, California Institute of Technology, 2000.

Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

Helen E Jones, Ian M Andolina, Bashir Ahmed, Stewart D Shipp, Jake TC Clements, Kenneth L Grieve, Javier Cudeiro, Thomas E Salt, and Adam M Sillito. Differential feedback modulation of center and surround mechanisms in parvocellular cells in the visual thalamus. *The Journal of Neuroscience*, 32(45):15946–15951, 2012.

Tilke Judd, Frø do Durand, and Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations A Benchmark of Computational Models of Saliency to Predict Human Fixations. 2012.

Nancy Kanwisher and Ewa Wojciulik. Visual attention: insights from brain imaging. *Nature Reviews Neuroscience*, 1(2):91–100, 2000.

Sabine Kastner and Leslie G Ungerleider. The neural basis of biased competition in human visual cortex. *Neuropsychologia*, 39(12):1263 – 1276, 2001. ISSN 0028-3932.

A.J. King, M.E. Hutchings, D.R. Moore, and C. Blakemore. Developmental plasticity in the visual and auditory representations in the mammalian superior colliculus. 1988.

Kurt Koffka. *Principles of Gestalt psychology.* Routledge, 2013.

B. Kolb and I.Q. Whishaw. Brain plasticity and behavior. *Annual review of psychology*, 49(1):43–64, 1998.

Z. Kourtzi, J.J. DiCarlo, et al. Learning and neural plasticity in visual object recognition. *Current opinion in neurobiology*, 16(2):152–158, 2006.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. *CoRR*, abs/1411.1045, 2014.

Koen Lamberts and Rob Goldstone. *Handbook of cognition*. Sage, 2004.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Fei Fei Li, Rufin VanRullen, Christof Koch, and Pietro Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14):9596–9601, 2002.

Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. 2013.

Nuo Li and James J DiCarlo. Unsupervised natural visual experience rapidly reshapes size invariant object representation in inferior temporal cortex. *Neuron*, 67(6):1062, 2010.

Yuewei Lin, Bin Fang, and Yuanyan Tang. A computational model for saliency maps by using local entropy. In *AAAI*, 2010.

Hongyu Li Lin Zhang, Zhongyi Gu. Sdsp: A novel saliency detection method by combining simple priors. In *IEEE International Conference on Image Processing (ICIP 2013)*, 2013.

S.G. Lomber, M.A. Meredith, and A. Kral. Cross-modal plasticity in specific auditory cortices underlies visual compensations in the deaf. *Nature neuroscience*, 13(11):1421–1427, 2010.

David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

W. Maass, R. Legenstein, and H. Markram. A new approach towards vision suggested by biologically realistic neural microcircuit models. In *Biologically Motivated Computer Vision*, pages 282–293. Springer, 2002.

Vijay Mahadevan and Nuno Vasconcelos. Biologically inspired object tracking using center-surround saliency mechanisms. 2013.

George R Mangun. Neural mechanisms of visual selective attention. *Psychophysiology*, 32(1):4–18, 1995.

F Mark, WC Barry, and AP Michael. Neuroscience: exploring the brain. *Nishimura Co., Ltd*, pages 644–658, 2007.

H. Markram. The blue brain project. *Nature Reviews Neuroscience*, 7(2):153–160, 2006.

D Marr. Vision, 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, 1982.

J.W. McClurkin, L.M. Optican, B.J. Richmond, et al. Cortical feedback increases visual information transmitted by monkey parvocellular lateral geniculate nucleus neurons. *Visual neuroscience*, 11:601–601, 1994.

Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. Yarp: yet another robot platform. *International Journal on Advanced Robotics Systems*, 3(1):43–48, 2006.

Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, pages 50–56. ACM, 2008.

Ethan Meyers and Lior Wolf. Using Biologically Inspired Features for Face Processing. *International Journal of Computer Vision*, 76(1):93–104, July 2007. ISSN 0920-5691.

Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. Active segmentation with fixation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 468–475. IEEE, 2009a.

Ajay Mishra, Yiannis Aloimonos, and Cornelia Fermuller. Active segmentation for robotics. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3133–3139. IEEE, 2009b.

Plinio Moreno and Manuel J Mar. A comparative study of local descriptors for object category recognition : SIFT vs HMAX. *Pattern Recognition*, (June):1–8, 2007.

Plinio Moreno, Manuel J Marín-Jiménez, Alexandre Bernardino, José Santos-Victor, and Nicolás Pérez de la Blanca. A comparative study of local descriptors for object category recognition: Sift vs hmax. In *Pattern Recognition and Image Analysis*, pages 515–522. Springer, 2007.

J. Mutch and D.G. G Lowe. Multiclass Object Recognition with Sparse, Localized Features. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, pages 11–18, 2006.

Jim Mutch and David G. Lowe. Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields. *International Journal of Computer Vision*, 80(1):45–57, January 2008. ISSN 0920-5691.

Jonathan J Nassi and Edward M Callaway. Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, 10(5):360–72, 2009.

Ugo Pattacini. *Modular cartesian controllers for humanoid robots: Design and implementation on the iCub*. PhD thesis, Ph. D. dissertation, RBCS, Italian Institute of Technology, Genova, 2011.

Marius V Peelen, Li Fei-Fei, and Sabine Kastner. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 460(7251):94–97, 2009.

Mary A Peterson. Object recognition processes can and do operate before figure-ground organization. *Current Directions in Psychological Science*, pages 105–111, 1994.

Rolf Pfeifer, Max Lungarella, and Fumiya Iida. Self-organization, embodiment, and biologically inspired robotics. *science*, 318(5853):1088–1093, 2007.

Tomaso Poggio, Sharat Chikkerur, and Others. Approximations in the HMAX Model. *Computational Complexity*, 2011.

Michael J Proulx and Howard E Egeth. Biased competition and visual search: the role of luminance and size contrast. *Psychological research*, 72(1):106–113, 2008.

G. Rachmuth, H.Z. Shouval, M.F. Bear, and C.S. Poon. A biophysically-based neuromorphic model of spike rate-and timing-dependent plasticity. *Proceedings of the National Academy of Sciences*, 108(49):E1266–E1274, 2011.

Babak Rasolzadeh, Mårten Björkman, Kai Hübner, and Danica Kragic. An active vision system for detecting, fixating and manipulating objects in the real world. *The International Journal of Robotics Research*, 29(2-3):133–154, 2010.

J.P. Rauschecker. Parallel processing in the auditory cortex of primates. *Audiology and Neurotology*, 3(2-3):86–103, 1998.

P Reinagel and R C Reid. Temporal coding of visual information in the thalamus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(14): 5392–400, July 2000. ISSN 0270-6474.

M Riesenhuber and T Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11), November 1999. ISSN 1097-6256.

M. Riesenhuber and T. Poggio. Models of object recognition. *Nature neuroscience*, 3: 1199–1204, 2000.

M. Riesenhuber, T. Poggio, et al. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2:1019–1025, 1999.

Ehud Rivlin and Héctor Rotstein. Control of a camera for active vision: Foveal vision, smooth tracking and saccade. *International Journal of Computer Vision*, 39(2):81–96, 2000.

M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–9, April 2007a. ISSN 0027-8424.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, 29(3):411–26, 2007b. ISSN 0162-8828.

T. Sharpee, N.C. Rust, and W. Bialek. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Computation*, 16(2):223–250, 2004.

Christian Siagian and Laurent Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(2):300–312, 2007.

Christian Siagian and Laurent Itti. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on*, 25(4):861–873, 2009.

A.M. Sillito, J. Cudeiro, and H.E. Jones. Always returning: feedback and sensory processing in visual cortex and thalamus. *TRENDS in Neurosciences*, 29(6):307–316, 2006.

O. Sporns, G. Tononi, G.M. Edelman, et al. Connectivity and complexity: the relationship between neuroanatomy and brain dynamics. *Neural Networks*, 13(8):909–922, 2000.

Yaoru Sun and Robert Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77 – 123, 2003. ISSN 0004-3702.

Nadia Tamayo and V Javier Traver. Entropy-based saliency computation in log-polar images. In *VISAPP (1)*, pages 501–506, 2008.

Hisashi Tanigawa, Haidong D Lu, and Anna W Roe. Functional organization for color and orientation in macaque V4. *Nature neuroscience*, pages 1–33, November 2010. ISSN 1546-1726.

Christian Theriault, Nicolas Thome, and Matthieu Cord. HMAX-S : Deep Scale Representation for biologically inspired Image Categorization. *Image (Rochester, N.Y.)*, pages 3–6, 2011.

Mick Thomure, Will Landecker, and Melanie Mitchell. Random prototypes in hierarchical models of vision. *Learning*, (1998):2010, 2010.

S. Thorpe, D. Fize, C. Marlot, et al. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

S.J. Thorpe and M. Fabre-Thorpe. Seeking categories in the brain. *Science*, 291(5502): 260–263, 2001.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

Martin J Tovee. How fast is the speed of thought ? 4(12):1125–1127, 1994.

Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

A. Ude and G. Cheng. Object recognition on humanoids with foveated vision. *4th IEEE/RAS International Conference on Humanoid Robots, 2004.*, 1(Humanoids):885–898, 2004.

A. Ude, C. Gaskett, and G. Cheng. Support vector machines and gabor kernels for object recognition on a humanoid with active foveated vision. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 1, pages 668–673. IEEE, 2004.

A. Ude, V. Wyart, L.H. Lin, and G. Cheng. Distributed visual attention on a humanoid robot. In *Proceedings of*, pages 381–386, 2005.

Aleš Ude, Damir Omrčen, and Gordon Cheng. Making object learning and recognition an active process. *International Journal of Humanoid Robotics*, 5(02):267–286, 2008a.

Aleš Ude, Damir Omrčen, and Gordon Cheng. Making Object Learning and Recognition an Active Process. *International Journal of Humanoid Robotics*, 05(02):267, 2008b. ISSN 0219-8436.

Sabine Kastner Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000.

B.M. Unit. Age-related changes in modular organization of human brain functional networks. *Neuroimage*, 44:715–723, 2009.

D C Van Essen, Charles H Anderson, Daniel J Felleman, Van Essen, and C David. Information processing in the primate visual system - An integrated systems perspective. *Science*, 255(5043):419–423, 1992.

Van Essen, D.C., J.W. Lewis, H.A. Drury, N. Hadjikhani, R.B.H. Tootell, M. Bakircioglu, and M.I. Miller. Mapping visual cortex in monkeys and humans using surface-based atlases. *Vision research*, 41(10-11):1359–1378, 2001.

Shaun P Vecera. Toward a biased competition account of object-based segregation and attention. *Brain and Mind*, 1(3):353–384, 2000.

W E Vinje and J L Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science (New York, N.Y.)*, 287(5456):1273–6, February 2000.

B. Voytek, M. Davis, E. Yago, F. Barceló, E.K. Vogel, and R.T. Knight. Dynamic neuroplasticity after human prefrontal cortex damage. *Neuron*, 68(3):401–408, 2010.

Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural networks : the official journal of the International Neural Network Society*, 19(9): 1395–407, November 2006. ISSN 0893-6080.

Dirk Walther, Ueli Rutishauser, Christof Koch, and Pietro Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image ...*, (November 2004):1–26, 2005.

Andrew B Watson. Probability summation over time. *Vision research*, 19(5):515–522, 1979.

Heiko Wersing and Edgar Körner. Learning optimized features for hierarchical models of invariant object recognition. *Neural computation*, 15(7):1559–88, July 2003. ISSN 0899-7667.

Robert J Wood. The first takeoff of a biologically inspired at-scale robotic insect. *Robotics, IEEE Transactions on*, 24(2):341–347, 2008.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5: 975–1005, 2004.

Steven Yantis. The neural basis of selective attention cortical sources and targets of attentional modulation. *Current Directions in Psychological Science*, 17(2):86–90, 2008.