DISSERTATION

Attention-driven Object Detection and Segmentation for Robotics

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Markus Vincze
E376
Institut für Automatisierungs- und Regelungstechnik

eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik

von

Dipl.-Ing. Ekaterina Potapova
geb. am 16.01.1988
Matr. Nr.: 1028784

Wien, im Oktober 2014

Ekaterina Potapova

Abstract

Vision is an essential part of any robotic system and plays an important role in such typical robotic tasks for domestic environments as searching and grasping objects in cluttered scenes. To be efficient, vision systems are required to provide fast object detection and segmentation mechanisms. In the past, attention mechanisms have been proposed to cope with the complexity of the real world by detecting and prioritizing the processing of objects of interest, and therefore guide the search and segmentation of objects. The goal of this thesis is to create an attention-based visual system, consisting of attention-based object detection and attention-driven object segmentation for a robot.

Many models of visual attention have been proposed and proven to be very useful in robotic applications. We address the problem of obtaining meaningful saliency measures based on such characteristics as the object height and surface orientations that appear to be qualitatively better than traditional saliency maps. Moreover, recently it has been shown in the literature that not only single visual features, based on color, orientation or curvature attract attention, but complete objects do. Symmetry is a feature of many man-made and also natural objects and has thus been identified as a candidate for attentional operators. However, not many techniques exist to date that exploit symmetry-based saliency. In this thesis, a novel symmetry-based saliency operator that works on 3D data and does not assume any object model is presented. We show that the proposed saliency maps are better suited for the task of object detection. Object detection was implemented by means of extracting fixation points from saliency maps. The evaluation in terms of the quality of fixation points showed that the proposed algorithms outperform current state-of-the-art saliency operators. The quality of attention points was defined in terms of their location within the object and the number of attended objects.

Segmentation of highly cluttered indoor scenes is a challenging task and traditional segmentation methods are often overwhelmed by the complexity of the scene and require a significant amount of processing time. To tackle this problem we propose to use attention-driven and incremental segmentation, where attention mechanisms are used to prioritize parts of the scene to be handled first. In this work, we combined a saliency operator based on 3D symmetry with three segmentation methods. The first one is based on clustering locally planar surface patches. The second method segments attended objects using an edge map based on color, depth and curvature within a probabilistic framework. We also proposed a third method, an incremental attention-driven mechanism, that outputs object hypotheses composed of parametric surface models. We evaluated our approaches on two publicly available datasets of cluttered indoor scenes containing man-made objects. We showed that the proposed methods outperform existing state-of-the-art attention-driven segmentation algorithms in terms of segmentation quality and computational performance.

This manuscript is best to read in color.

Acknowledgement

This work would not have been possible without the help and support from my family, friends and colleagues. I would like to sincerely thank my supervisor, Prof. Markus Vincze, for giving me such a wonderful opportunity and freedom to do my research under his supervision at Vienna University of Technology, Automation and Control Institute, Vision for Robotics Group. I would also like to thank my second supervisor, Dr. Simone Frintrop, for taking time in reviewing my thesis, as well as the short, but amazing collaboration I experienced working with her team. My special gratitude goes to my mentor, Dr. Michael Zillich, for his time, help, support, ideas and continuous productive discussions about my work.

Many thanks go to my former and present colleagues – it was fun working with you guys! Special gratitude goes to Paloma de la Puente for taking her time and reading my thesis and to Karthik Mahesh Varadarajan for his continuous effort in helping me to improve my English skills over the years.

The research leading to the results in this thesis has received funding from the European Community, Seventh Framework Programme (FP7/2007-2013), under grant agreements No. 610532 (SQUIRREL) and No. 215821 (GRASP) and from the Austrian Science Foundation (FWF) under grant agreement No. TRP 139-N23 (InSitu).

Finally I want to express my love and gratitude to my family, especially to my mother, for her unconditional love, support and encouragement. Without her I would not have been able to go so far in my academic and professional career. I also want to thank my better half, Sebastian for his support, optimism and continuous belief in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last 20 years, robotics has become increasingly widespread, aiming to assist, protect and simplify the everyday life of people. The first robots were built for industry as a tool to automate production. Nowadays, robots are intended to help and serve humans. The growing number of robots, and areas they are used in, raise multiple questions and problems for researchers developing robotic systems.

Robots find increasing use in domestic environments and are currently being introduces to perform a wide variety of assistance tasks. Examples of scenarios for domestic robots reach from assistance in the kitchen, such as sorting dishes and loading a dishwasher, to searching for and bringing a specific object from the living room to the human. When placed in a domestic environment, robots face challenges that are significantly different from laboratory conditions. In particular, robots in such environments have to deal with a significant amount of clutter and increased world complexity. This complexity is partly caused by the variety of objects a robot has to deal with. These objects can be multi-colored, textured, partly occluded, etc. Some objects do not even have a corresponding a-priori model available for the system. A robot is often issued a task that requires not only to find such objects, but also interact with them. Such tasks include, but are not limited to:

- *searching for a known object*: “Bring my cup!”
- *searching for an unknown object*: “Bring me the yellow book over there!”
- *searching for general objects*: “Clean up the table!”

For example, let us have a look at the robot shown in Fig. 1.1. It was specifically designed to be placed in households to help people. One simple task the robot can be asked to perform by a human is “Bring my cup!”. To fulfill this task the robot hast to (1) find the cup, (2) grasp it, and (3) bring it to the user. To successfully complete the first step – finding the cup – the robot has to be able to search and localize the required object. In other words the robot has to detect the object. By definition, “object detection” is a technique in computer vision and image processing that deals with detecting, in digital images and videos, instances of objects of a certain or a generic class. There exist two main groups of approaches for object detection: top-down and bottom-up.

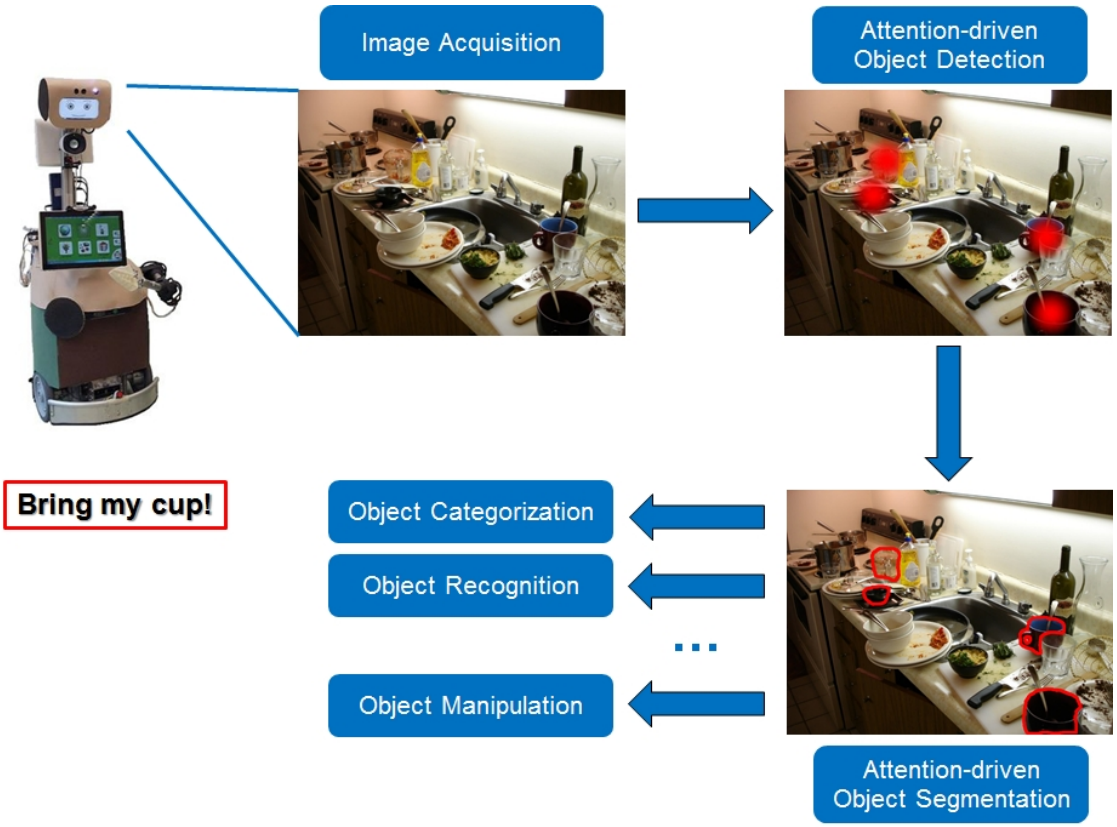


Figure 1.1: A general pipeline of the proposed attention-driven robot perception strategy when the user demands the robot: "Bring my cup!"

Top-down approaches for object detection typically include sliding window detectors [Viola 04, Alexe 12, Manén 13, Rahtu 11, Feng 11]. The sliding window detectors are conceptually simple: independently classify all image patches as being object or non-object. Sliding window detection is a very popular concept for object detection in computer vision. Such detectors usually operate on color images and detect object candidates based on a set of matched image features combined with machine learning. Some of these features were trained to detect specific categories of objects, e.g. faces such as in [Viola 04]. Others were specifically designed to detect generic objects [Alexe 12, Rahtu 11, Feng 11]. Because all image patches have to be verified against all known object categories, such approaches are usually computationally inefficient. Moreover, the typical number of generated object proposals is usually very high, while the detection quality is low. This means that the object proposals have to be verified and filtered at later stages to select only those that really represent objects. Additionally, each specific object type should have its own detector, or in case of generic object detectors each object candidate should be run through the recognition process. The size of detected objects is tightly linked to the size of the sliding window. Therefore more image patches have to be checked when the object size is unknown.

Bottom-up approaches for object detection are based on the full scene segmentation. Scene segmentation [Arbelaez 12, Comaniciu 02, Felzenszwalb 04] is the process of partitioning the scene into object candidates. Then those segmented object candidates are

checked against known object models. Segmentation itself is an ill-defined problem, because for different tasks correct segmentation of the same scene can be different. The complexity and required computational resources increase, when the scene is cluttered, multi-textured and multi-colored. Object candidates have to be run through the recognition process after segmentation, which also slows down the speed of the object detection.

Taking into consideration that multiple modules, such as navigation and localization, categorization and recognition, speech understanding, human-robot interaction and many others, have to run on the robot at the same time, traditional object detection approaches are not suitable for use in robotics.

1.1 Problem Statement

Searching exhaustively through every image for a required object, as is done in traditional object detection approaches, is very computationally expensive. Moreover, with such an approach, a lot of false object candidates are detected. The reason for multiple false positives lies in the way those approaches are designed and tested. Usually they work very well on the databases they were trained on. However, when they are used in real world environments with changing illumination, pose and appearance, and previously unseen objects, they may result in identifying non-objects as target objects. Another problem with using classical approaches is that they typically output a lot of object proposals and those proposals have to be checked against being the object of interest. This is a very costly operation. What we would like to have in robotics is a small number of good object proposals.

The inspiration to overcome those problems can be drawn from underlying processes in the human visual perception system, which solves tasks such as object detection so effortlessly. One of the capabilities that humans possess making us efficient in processing visual stimulus is called the visual attention system. The human visual attention system is an efficient mechanism for prioritizing visual information about the surroundings and selecting important regions. Attention, as well as computational models of attention producing saliency maps, gain significance especially in cluttered scenarios and are essential in dealing with the uncertainty, as well as increased complexity, coming from the real world. Potential objects are detected as areas of highest attention.

Research on human visual attention is traditionally done using color images [Itti 98, Frintrop 05a]. Computational visual attention models output saliency maps. Object detection is usually implemented as extraction of fixation points, also called attention points, from the saliency map. One well-known extraction algorithms is the Winner-Take-All neural network of [Lee 99], which outputs locations of locally highest saliency values.

However, there are several problems with using standard color-based saliency maps and extraction algorithms when applied to robotics. For example, color-based saliency maps detect regions of high contrast with respect to the surroundings. In cluttered and multi-colored scenes, such an approach results in a huge number of false detections. Moreover, for robotic applications, meaningful saliency measures are required to be connected to the task a robot has to execute. For example, salient objects for the task of grasping do not necessarily correspond to regions of high color contrast.

Another problem lies in the re-detection of the same object several times. One of the reasons for that lies in the nature of saliency maps, that were not designed to detect separate objects. Another reason for re-detection is because of the used extraction strategies. These are typically designed to detect locations of the highest local saliency values. If a big part of an object is salient, the extraction strategy will detect several fixation points on the same object.

Recently, it has been shown in the literature on visual attention that not only single visual features, such as color, orientation or curvature attract attention, but also complete objects do [Einhäuser 08]. For example, symmetry is a feature of many man-made and natural objects, and has thus been identified as a candidate for attentional operators. However, not many techniques exist to date that exploit symmetry-based saliency. So far these techniques work mainly on 2D data. Furthermore, methods, which work on 3D data, assume complete object views, which are not available when robot is browsing a search space. This limits the usage of such operators as bottom-up attentional operators working on RGB-D images, which only provide partial views of objects.

In robotics, detection of object for a specific task, or generic objects with minimal number of repetitive detections, is of high importance. Moreover, developing a new extraction strategy, that results in an increased number of correctly detected objects is also necessary to improve the overall performance.

Detected objects are then used in further tasks, such as recognition or manipulation. For many of those tasks the shape of the object is required, e.g. for grasping, to calculate grasping points. Therefore, after an object is detected it should be segmented.

Common approaches in computer vision segment the complete scene, as described before. However, in robotic tasks it is sufficient to segment only the object of interest. For example, for the task of grasping it is enough to detect and grasp the first suitable object and then observe the scene again and make a subsequent decision step. What we essentially want is a fast and reliable segmentation only of the detected object. Moreover, in cluttered environments it is more preferable to concentrate on the local properties to separate a specific object from the background and other objects. Segmentation methods that successfully solve the above given tasks are attention-driven segmentation techniques [Mishra 11, Kootstra 10a]. Attention-driven segmentation approaches take as input a fixation point, belonging to the object. Starting from this point they identify parts of the scene that belong to the required object. This class of segmentation methods are implemented in a very efficient manner. However, even current state-of-the-art attention-driven segmentation approaches have some significant drawbacks. The most important of which is imprecise segmentation – in the form of over or under segmentation. Ideally, the object should be segmented precisely, because the precision of the segment influences the quality of recognition and manipulation.

1.2 Proposed Approach

To solve the above described problems concerned with object detection and segmentation in robotics, we propose to view detection and segmentation as a fundamentally attention-

driven problem.

The system consists of two parts: attention-driven object detection and attention-driven object segmentation. We argue that this attention-driven perception strategy is a way to successfully resolve increasing complexity which appears when robots have to deal with complicated real world situations. Each of the tasks of object detection and further segmentation is a very challenging research problem on its own.

An example of how the proposed system can be used in real world situations is explained using scenario shown in Fig. 1.1, where a robot is asked to "Bring my cup!". At the first stage, the proposed visual system detects the potential locations of of all hypothetical proto-objects using attention. Those regions may potentially represent objects, when the proper attention mechanism is used. Such locations are shown as red blobs in the image. We propose to detect those locations as fixation points representing the regions of highest attention in the scene. At the second stage, given fixation points, that belong to objects, attention-driven segmentation is performed. Segmented objects are indicated with bounding boxes in the picture. The segmented objects can be sent to other modules, such as recognition, manipulation, etc.

1.3 Contributions & Publications

The main novelties and contributions presented in this thesis with respect to attention-driven object detection for robotics are the following:

- To investigate the importance of 3D visual information for humans we conducted a deep and comprehensive analysis of findings and conclusions from eye-tracking experiments in 3D. Based on the analysis we conclude that 3D information plays an important role in human vision. This means that ways to integrate it into the robotic vision system should be investigated. To our knowledge this is the first systematic analysis of eye-tracking experiments performed in 3D environments.

- To understand the usability for robotics of existing attentional models, we conducted their systematic comparison. Attentional models are compared with each other in terms of the scope they were designed for and principles they were built on. To the best of our knowledge, this is the first systematic comparative analysis of existing attentional models that involve 3D information.

- To detect objects for robotic tasks, such as grasping, we propose a set of saliency cues, based on such properties as height and surface orientation. We show that fusion of those saliency cues into saliency maps, while appearing to be similar to traditional saliency maps, represents better object detection abilities. Parts of this research have been published in the Proceedings of the International Conference on Computer Vision Systems (ICVS) [Potapova 11].

- To detect generic objects in cluttered scenes, we propose a novel 3D symmetry-based saliency operator that works on partial object views. To improve object detection

quality we designed a new extraction strategy specifically for the 3D symmetry-based saliency operator. Results show increase in detection quality of generic objects, compared to existing state-of-the-art saliency maps and extraction strategies. Parts of this research have been published in the Proceedings of the Asian Conference on Computer Vision (ACCV) [Potapova 12b].

- To show benefits of different multi-level strategies, combination methods and normalization operators, that are used to create saliency maps, for the task of object detection, we perform exhaustive evaluation of the proposed saliency cues against state-of-the-art saliency operators. Our detailed analysis shows that more complicated, and therefore more computationally expensive, methods do not improve detection performance.

The main novelties and contributions presented in this thesis with respect to attention-driven object segmentation for robotics are the following:

- To segment detected objects in highly cluttered scenes, we propose two novel attention-driven segmentation algorithms. The first algorithm segments objects using an edge map based on color, depth and curvature within a probabilistic framework. The second segmentation algorithm is based on clustering locally planar surface patches. The evaluation on indoor table-top scenes containing man-made objects clustered in piles and dumped in a box shows that proposed algorithms lead to a significant improvement compared to state-of-the-art methods of active segmentation. Parts of this research have been published in the Proceedings of the International Conference on Pattern Recognition (ICPR) [Potapova 12a] and in the Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) [Potapova 14b].

- To extend attention-driven segmentation to be able to segment as many objects as necessary, in comparison to only one object indicated by the baseline attention-driven segmentation, we develop an incremental segmentation method. Incremental segmentation uses attentional mechanisms to detect and prioritize processing of objects of interest. This basically means that segmentation starts from the object of the highest importance and incrementally explores the scenes based on the importance of each region. Eventually, given enough time, incremental segmentation segments the complete scene. Evaluation on cluttered indoor scenes shows that the proposed method outperforms existing methods of attention-driven segmentation in terms of segmentation quality and computational performance. Parts of this research have been published in the Proceedings of the IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS) [Potapova 14a].

1.4 Outline

This thesis consists of two main chapters. In Chapter 2, we give an exhaustive overview of existing attention operators and attention-based object detection strategies. The contributions of this chapter are: (1) a deep and comprehensive analysis of existing eye-tracking

experiments in 3D and attention models for RGB-D data from the perspective of a robotic visual system; (2) the development of a set of novel attentional operators based on RGB-D data that result in saliency maps; (3) an investigation and evaluation of different combination and normalization strategies for plausible saliency cues; (4) the development of novel algorithms, that reliably detect objects in indoor cluttered environments, by means of extracting fixation points, or so-called attention points, from saliency maps; (5) an exhaustive evaluation of several existing attention point extraction strategies against different types of state-of-the-art saliency maps and proposed approaches in terms of uniqueness and proximity of fixations to the center of the detected object. Results show that our proposed 3D-based saliency operators and extraction strategy reflect the notion of objectness better than the currently existing state-of-art 2D-based and 3D-based saliency operators.

In Chapter 3, we continue with a discussion of attention-driven segmentation and its benefits when applied to real world robotic tasks. Three novel methods for attention-driven segmentation for cluttered table scenes are proposed: (1) attention-driven segmentation with object detection based on attention points from 3D symmetry saliency maps [Potapova 12b], (2) attention-driven segmentation based on probabilistic edges, and (3) incremental attention-driven segmentation based on complete-scene segmentation proposed by Richtsfeld *et al.* [Richtsfeld 12]. All proposed methods are evaluated against current state-of-the-art algorithms and it is shown that obtained results outperform existing approaches.

The thesis is concluded with Chapter 4 and possible directions of future research are discussed.

Chapter 2

Attention-driven Object Detection

Human visual attention is a cognitive phenomenon serving as a mediator in the selective process of identifying valuable parts in the environment. The amount of visual information received by the retina in every single moment of time is too large to process and visual attention is a natural mechanism that prioritizes visual space by concentrating on specific parts while ignoring others. Attention has become an important topic in the computer vision and robotics communities, since the attention paradigm was proven to be extremely useful when dealing with complex real world scenarios. The purpose of visual attention is seen to be as reduction of the scene complexity and hence resulting in savings in computational resources. First theories of visual attention appeared in the second half of the 20th century and tried to explain human visual behavior by means of simplified psychological experiments [Treisman 80, Koch 85, Wolfe 89, Tsotsos 90]. Later, computer vision and image processing started exploiting visual attention to the benefit of artificial vision systems: object recognition [Walther 04], image retargeting [Achanta 09b], image compression [Yu 09], scene classification [Borji 11], visual tracking [Mahadevan 12], and many others, showed how to deploy visual attention to increase the performance and the quality of results. One of the first papers that argued about and demonstrated the significance of visual attention for robotic systems was published by Aloimonos *et al.* [Aloimonos 88]. Almost 20 years later, Tsotsos and Shubina [Tsotsos 07] discussed importance of attention for robotics and demonstrated how object search in a 3D space greatly benefits by integrating it. One of the landmark examples that show how attention can be efficiently utilized in robotics is active segmentation [Mishra 09, Björkman 10], where visual attention is used as a first step to detect objects for tasks such as manipulation.

Research on attention distinguishes between two mechanisms of visual attention: bottom-up and top-down [Yarbus 67]. Bottom-up visual attention is a signal-driven mechanism that operates on raw sensory data and involuntarily shifts attention towards salient visual features of the observed scene. Top-down visual attention [Navalpakkam 06, Frintrop 05a] is a voluntary mechanism that strongly depends on the viewing task and prior information about the environment. It was shown that these two mechanisms strongly interact with each other and affect human visual behavior [Wolfe 03]. The majority of computational models for visual attention were developed to explain bottom-up attentional mechanism [Itti 98, Harel 06, Hou 07, Achanta 08]. These computational models produce saliency

maps, where higher the saliency value of the region, the more attention it attracts.

For bottom-up saliency there are two prevalent models: contrast-based and information-based. Contrast-based models calculate saliency by analyzing spatial contrast. These models primarily concentrate on low-level visual features, such as luminance, color, and orientation. An example of contrast-based models is the model proposed by Itti *et al.* [Itti 98], which exploits a center-surround mechanism to define saliency across scales and was inspired by the neural mechanism. The model has established itself as the de-facto benchmark for saliency detection and is often used for comparison in literature. Information-based models calculate saliency in terms of information maximization or frequency domain analysis [Hou 07, Bruce 09, Zhang 08]. These models concentrate on detecting salient regions based on their rareness and uniqueness in the scene and possibly include the notion of surprise [Itti 06].

Einhäuser *et al.* [Einhäuser 08] have shown that not only single popping out salient features or unique scene regions attract visual attention, but also complete objects. Following the idea of object-based saliency, Kootstra *et al.* [Kootstra 08] proposed to use symmetry as a feature to identify potentially attractive objects in the scene. The authors argued that for symmetrical objects, a symmetry-based saliency map corresponds better to human visual attention than contrast-based saliency maps.

Almost for two decades, computational models of visual attention were concerned only with processing color images. However, the increasing amount of 3DTV displays, appearance of cheap 3D sensors such as Kinect and Time-of-Flight cameras forced the community to investigate how depth information may influence human visual perception. Over the last 10 years a number of publications proved that the presence of depth information changes the behavior of the human attentional mechanism [Jansen 09, Häkkinen 10, Huynh-Thu 11]. This discovery is of particular interest for robotics, where depth information plays a crucial role for navigation and localization and typically robots are equipped with RGB-D sensors. Even before the first eye-tracking experiments in 3D were reported, the robotics community has started exploiting depth as an additional channel for saliency map computation [Frintrop 05b, Maki 96, Ouerhani 00, Akman 10].

Robotic systems have multiple tasks to solve in parallel, including localization and navigation, manipulation, segmentation and recognition, and many others. Therefore, vision for a robot is the only one fraction of a larger system. Because of several concurrent processes running on the robot, the size of visual search space, and such essential requirements as robustness and real-time performance, integrating visual attention is a necessary prerequisite for any successfully operating vision-based robotic system. Integrating attention as a part of the vision system means that instead of processing the complete scene, the robot will concentrate its resources only on the most attractive and relevant regions in the scene.

In this thesis, the focus is on investigating bottom-up computational models of visual attention that involve depth processing with application to robotic visual systems. We investigate attention for the task of object detection, where objects can appear in various locations and configurations, partly occluded and surrounded by clutter. To solve the detection problem, we propose a set of attentional operators based on RGB-D data. Given a set of plausible saliency cues, we also aim at investigating strategies for combining these cues.

As will be described in the related work section (see 2.1) there are several ways to evaluate the performance of a visual attention model. In this thesis, the focus is on the goal of reliably detecting objects in indoor cluttered environments, and therefore we propose to extract fixation points [Lee 99], or so-called attention points, from saliency maps. We evaluate several existing attention point extraction strategies against different types of state-of-the-art saliency maps in terms of uniqueness and proximity of fixations to the center of the detected object. We also propose a novel approach to extract attention points from saliency maps and show that it outperforms existing strategies. Our results show that our proposed 3D-based saliency operators reflect the notion of objectness better than the currently existing state-of-art 2D-based and 3D-based saliency operators.

The chapter is structured as follows: In the next section 2.1, we describe related work on saliency map calculation, subsequently, attention operators and attention point extraction strategies are discussed in detail (2.3–2.5). The following evaluation in section 2.6 shows comparison of different attention operators and attention point extraction strategies.

2.1 Related Work

In this Section, we focus on the description of existing algorithms for bottom-up saliency computation. We start with an overview of the best known saliency algorithms for color images and continue with a detailed description of eye-tracking experiments in 3D and existing algorithms developed for RGB-D data. The purpose of this section is to give the reader an understanding of the field of computational visual attention, its applicability and limitations for robotic vision systems in real world scenarios.

2.1.1 Attention Operators for Color Images

In the following Section, we give an overview of the best known bottom-up visual attention models to generate saliency maps from color images that are used for further evaluation. We also discuss the scope of their usability and relation to robotics. A detailed overview of the state-of-the-art saliency algorithms for color images can be found in [Borji 13].

In 1998 Itti *et al.* [Itti 98] published their influential paper on the model of saliency-based visual attention. Over the next 15 years, this work has been cited more than 5000 times and still serves as a comparison model for many saliency algorithms. The model of saliency-based visual attention was inspired by the neural network similar to the one found in human brains and exploits ideas of the feature integration theory proposed by Treisman and Gelade [Treisman 80]. The core idea of the model is to hierarchically decompose the image based on three low-level visual features (intensity, color and orientation) and to compute the center-surround difference. The center-surround mechanism responds to image differences between a small central region and a broader concentric surround region. At a later stage, these feature maps are fused together to obtain a saliency map. Over the next several years, the proposed algorithm was extended with different normalization operators and features [Itti 99, Itti 00, Itti 01]. Later, Frintrop [Frintrop 06a] proposed to use center-off as well as center-on mechanism to detect interesting locations in images and showed how beneficial it is for a robotic vision system to use both.

In 2006, Harel *et al.* [Harel 06] described a simple and biologically plausible model for bottom-up saliency. The introduced model computes activation maps, similar to [Itti 98], formed on certain feature channels, such as color and orientation, which are then combined into the master saliency map. The main difference to [Itti 98] is that instead of using a hierarchical structure, a fully-connected directed graph where weights of the edges depend on the similarity between pixels they connect, is constructed and used to calculate activation maps. The graph is treated as a Markov chain and the activation map is considered to be an equilibrium state. The Markov chain is further used to concentrate intensity of activation maps in order to normalize them.

In 2007, Hou *et al.* [Hou 07] proposed a simple and efficient method independent of features, categories and other forms of prior knowledge to detect saliencies in the image. This was one of the first algorithms that proposed to detect saliency in the spectral domain using Fourier transform and argued that visual saliency appears in regions that are globally different to the image content. As a result, the algorithm analyzes the log-spectrum of an input image and calculates saliency as the spectral residual.

In 2009 Bruce *et al.* [Bruce 09] proposed a model of bottom-up attention based on the principle of maximizing information sampled from the scene. The model defines saliency by quantifying the self-information of each local image patch. Independent Component Analysis is performed on a large number of patches sampled from an image set to determine a suitable basis. The probability of observing a specific patch is evaluated by independently considering the likelihood of each corresponding basis coefficient. Shannon's self-information measure, applied to the joint likelihood of patch statistics, provides an appropriate transformation between probability and the degree of information inherent in the local statistics. Therefore, saliency is determined as the self-information of each local image patch.

In 2010 Kootstra *et al.* [Kootstra 10b] proposed to use symmetry, one of the Gestalt principles for figure-ground segregation, to calculate saliency maps. A symmetry-based saliency map [Kootstra 10b] is built upon the local context free symmetry operator of Reisfeld *et al.* [Reisfeld 95] and is extended to a multi-scale model similar to [Itti 98]. The amount of local symmetry at a point is calculated as the sum of similarity measures between pixel pairs in the symmetric kernel centered at the point. The similarity measure takes into consideration gradient directions and magnitudes, as well as distances between points.

The idea behind the algorithm is that saliency appears not only based on independent features, but is rather located on objects, as it was discussed in [Scholl 01, Einhäuser 08]. This finding is supported by experiments conducted by Vishwanath and Kowler [Vishwanath 04] who tested the sensitivity of saccadic localization to perceived 3D structure using computer-generated perspective images of simulated 3D objects. They found out that saccades made to 2D perspective images of a capsule-shaped 3D object landed near the 2D Center of Gravity (COG), while saccades made to 3D objects were shifted to the 3D COG. They also compared landing positions for 3D shapes that had identical 2D images, but different 3D interpretations and found that different participants landed either closer to 2D or 3D COG. They have also found that when more 3D cues are involved there was a displacement towards 3D COG. The pattern results obtained with the 3D shape sug-

gests that landing positions were based on pooling of information across the shape, rather than selection of a distinct position. In the light of these findings building saliency models based on object properties such as 3D symmetry is completely justified.

All of the described algorithms were developed with the purpose of mimicking human visual attention and were evaluated on natural scenes against human eye-tracking data. While such models can be and are successfully applied in the field of robotic vision, as described in [Begum 11], the primary goal of attention system for a robot is not to act in a way similar to humans, but rather to detect important and relevant parts of a scene in order to make the robot operate smoothly and efficiently. Many robots operate in human-made environments, where they face typical indoor scenes with variety of occlusions, colors and textures. These occlusions, colors and textures will eventually attract attention according to the color-based saliency models, while the goal is to detect object instances instead. This limits the application of color-based attention algorithms that were developed to detect saliencies in natural outdoor scenes. In our evaluation section, we will discuss performance of the above described algorithms in indoor environments and show examples where they do not show as good results as saliency operators proposed in this thesis.

2.1.2 Eye-tracking in 3D

As it was mentioned earlier the majority of computational attention algorithms were developed for color images and evaluated against human eye-tracking data. The increase in the popularity of stereoscopic displays and 3DTV in recent years has boosted research on attention and eye-tracking in 3D. Multiple researchers have reported that depth matters when it comes to scene exploration ([Hügli 05, Wexler 08, Jansen 09, Ramasamy 09, Häkkinen 10, Wismeijer 10, Liu 10, Huynh-Thu 11, Wang 12, Lang 12b, Gautier 12, Wang 12] and [Khaustova 13]). Robotics has a long history of using laser scanners to obtain depth to be used in such tasks as navigation [Biswas 12, Cunha 11, González-Jiménez 13]. With the emergence of cheap and ready to use 3D sensors such as the Microsoft Kinect, use of depth has increased for tasks of segmentation [Ückermann 12], object recognition [Aldoma 13], human tracking and pose recognition [Shotton 11], etc.

This Section is structured as follows: first, we discuss related work on eye-tracking experiments in 3D and explain how depth information influences human visual behavior; then we present insights on how those findings can be utilized for a robotic attention system. The overview of all eye-tracking experiments and their findings that involved depth information is presented in Table 2.1.

First eye-tracking experiments in 3D date back to the first decade of the 21st century. In 2005, Jost *et. al* [Jost 05] used random dot stereograms to investigate effects of "pure" depth information. Each random dot stereogram consisted of 4 objects, which were drawn with different disparities. It was found that the density of fixations was clearly higher in regions, where the disparity values were higher. It suggested that closer objects attract earlier fixations and therefore are often more salient than those situated further away.

In 2008, Wexler and Ouarti [Wexler 08] investigated the effect of 3D plane orientation on spontaneous eye movements, as well as, various aspects of 3D scenes that potentially affect visual behavior. In the experiments, subjects were looking at planar surfaces inclined

		PAPER	[Jost 05]	[Wexler 08]	[Jansen 09]	[Ramasamy 09]	[Häkkinen 10]	[Wismeijer 10]	[Liu 10]	[Huynh-Thu 11]	[Wang 12]	[Lang 12b]	[Gautier 12]	[Khaustova 13]	[Desingh 13]
DATA	STATIC	STEREOGRAMS	●	–	–	–	–	–	–	–	●	–	–	–	–
		PLANES	–	●	–	–	–	●	–	–	–	–	–	–	–
		LANDSCAPES	–	–	●	–	–	–	●	–	–	–	●[11]	–	–
		INDOOR	–	–	–	–	–	–	–	–	–	●	–	●	●[13]
	VIDEO		–	–	–	●	●	–	–	●	–	–	–	–	–
FIXATIONS	DURATION		–	–	↗2D[3]	–	–	–	–	↗2D	–	–	–	–	–
	DISTANCE[1]		↘3D	–	↘3D[4]	–	–	–	–	–	↘3D	↘3D	↘3D	–	↘3D
	AMMOUNT		–	–	↗3D	–	–	–	–	–	–	–	–	–	–
	SPATIAL EXTENT		–	–	↗3D	↗2D[7]	↗3D	–	–	↗3D[9]	–	↗3D	–	–	–
	FREQUENCY		–	–	–	–	–	–	–	↗2D[10]	–	–	–	–	–
SACCADES	VELOCITY		–	–	–	–	–	–	–	↗3D	–	–	–	–	–
	LENGTH		–	–	↘3D[5]	–	–	–	–	–	–	–	–	↘3D	–
SALIENT FEATURES	DEPTH GRADIENT		–	●	●[3]	–	–	●	●[8]	–	–	–	–	–	–
	MEAN LUMINANCE		–	–	3D[6]	–	–	–	–	–	–	–	–	–	–
	LUMINANCE CONTRAST		–	–	2D/3D[6]	–	–	–	●	–	–	–	–	–	–
	TEXTURE CONTRAST		–	–	2D/3D[6]	–	–	–	–	–	–	–	–	–	–
	LUMINANCE GRADIENT		–	–	–	–	–	–	●	–	–	–	–	–	–
	DEPTH CONTRAST		–	–	–	–	–	–	●[8]	–	–	–	–	–	–
	COGNITIVE INFORM.[2]		–	–	–	–	–	–	–	●	–	●	–	–	–
	CENTER BIAS		–	–	–	–	–	–	–		–		●[12]	●[1]	●
	OBJECTS		–	●	–	–	–	–	●	–	–	–	–	●	●[14]

Notes:

[1] only in the beginning of viewing

[2] such as text and faces

[3] except for pink and white noise images

[4] also present in 2D

[5] over time

[6] only in natural images

[7] analyzed only one video containing a long deep hallway

[8] lower values were more salient

[9] only for videos with a single scene and static camera view

[10] sometimes similar in 2D and 3D

[11] used eye-tracking data from [Jansen 09]

[12] for 3D is much lower with time

[13] region-based method similar to [Li 13] instead of eye-tracking

[14] for objects in the focus

Table 2.1: The table shows comparison of different 3D eye-tracking strategies.

in depth simulated on a computer screen. Three types of inclined surfaces (grid, texture and dots) were used. Spontaneous exploratory eye movements of the participants were recorded during the visual task. As part of the experiment, subjects were required to report perceived 3D orientation of the plane. Experimental results revealed that three-dimensional surface orientation has an impact on human visual, because eye movements in the beginning show a tendency to follow the axis of the 3D plane tilt and surface depth gradients, while binocular disparity dominates the vergence. Interestingly, it was also found that this

effect holds in with and without given task. Authors speculated on two hypotheses to explain the discovered effect: (1) the depth gradient hypothesis, meaning that gaze moves back and forth along the depth gradient axis; and (2) the 3D-shape hypothesis, meaning that gaze moves along the principal axis of the perceived three-dimensional object.

In 2009, Jansen *et. al* [Jansen 09] studied how disparity information influences fixations and saccades. Visual stimuli with natural landscapes without any man-made objects were selected to understand what role disparity plays in a bottom-up visual attention. The analysis was conducted on the data from the left eye. The data from the left eye was used because 2D images were created using two copies of the left image. They investigated the saliency of several image features: mean luminance, luminance contrast, texture contrast, mean disparity (distance), and disparity contrast (depth discontinuity). It was figured out that disparity influences basic eye movements. As a result (1) the number of fixations increases, (2) the fixation duration decreases with time (only for pink and white noise), (3) the saccade length shortens over time, and (4) the spatial extent of exploration increases. The duration of fixations was not prolonged when specific tasks were given. The mean disparity changed viewing behavior in the beginning, meaning closer locations were observed at first. Later, attention was shifted to more distant locations. However, this behavior was also observed in 2D natural images. This means that neither the disparity nor the center-bias cannot explain eye-tracking results. In 3D noise images participants tend to look more at depth discontinuities rather than planar surfaces. In general, the mean luminance, the luminance contrast, and the texture contrast showed comparable saliency both in 2D and 3D. Texture contrast was a salient feature in natural, but not in noise images. Only in 3D natural images the mean luminance was correlated with fixations. The luminance contrast appeared to be salient in 2D and 3D natural images in the beginning of viewing. Authors could have concluded from the results that (1) 2D visual feature are important to design a 3D visual attention model 2) existing 2D visual attention models could be adapted to 3D visual attention.

In 2009, Ramasamy *et. al* [Ramasamy 09] presented work about using eye-tracking for stereoscopic film-making. The goal was to detect elements distracting people from the movie. The fixation patterns of one scene were analyzed to identify regions of interest. In a specific scene, it was found that in the non-stereoscopic version, the gaze data was more spread out across the frame and is not as highly focused into a long deep hallway at the other end of the doorway. Rephrasing, obtained results show that fixation spread could be potentially more concentrated while viewing 3D videos (for example for video filming long deep hallway). This opposes conclusions made by Jansen *et. al* [Jansen 09] and Häkkinen *et. al* [Häkkinen 10].

In 2010 Häkkinen, *et. al* [Häkkinen 10] studied the how fixation patterns differ from each other for 2D and 3D videos. Four short videos were selected to perform the study. Eye-tracking results have been analyzed by examining fixations in the image regardless the gaze depth. The results showed that fixations for 3D videos are more widely distributed, which supports findings by Jansen *et. al* [Jansen 09] and contradicts the conclusion made by Ramasamy *et. al* [Ramasamy 09]. They also reported that in 3D observers did not only pay attention to main characters of the video, but fixations were also on other objects. This shows that depth gives humans supplementary information, which results in new salient

areas. Potentially a saliency map from depth exists, as well as a mechanism that integrates 2D and 3D saliency information. Moreover, because in the study, participants were asked to compare 2D vs. 3D versions of videos, it can be argued that experiments had some top-down influence.

Following the work of Wexler and Ouarti [Wexler 08], Wismeijer *et. al* [Wismeijer 10] studied the correlation between perception and spontaneous eye movements and investigated if saccades are correlated with single depth cues or with a their combination. Observers viewed inclined surfaces where monocular perspective and binocular disparity showed different plane orientations and the degree of the conflict was varying from small to large. Three different response measures were used: (1) explicit reports of perceived surface tilt, (2) the directions of spontaneous saccades when scanning the stimulus, and (3) the orientation of the plane defined by the vergence of spontaneous eye movements. They showed that spontaneous saccade directions tend to follow depth gradients (surface tilts) and have the same combination pattern as perceived surface orientation. For small conflicts the combination is a weighted linear, while for large conflicts there is a cue dominance. Authors could have concluded from the experiments that binocular disparity dominates vergence. The results implied that the interocular distance should be compensated by the local disparity value.

In 2010, Liu *et. al* [Liu 10] studied fixated positions with respect to visual features. Stereo images with natural scenes were used in the experiment. The focus of the study was on examining visual features extracted at random fixation locations while looking at 3D static images. They used (1) the luminance gradient, (2) the disparity gradient, (3) the luminance contrast and (4) the disparity contrast maps and tried to figure out if human fixations correlate with those maps. They showed that the luminance gradient and the luminance contrast were overall had higher values at fixated locations. However, the disparity contrast and the disparity gradient at fixations were in generally lower than at random locations. The reported results are opposite to the results of Jansen *et. al* [Jansen 09], where it was found that humans look more consistently at depth discontinuities instead of planar surfaces. Liu *et. al*
[Liu 10] suggested that the binocular visual system, if not directed otherwise by a top-down task, fixates on areas that help to simplify the process of depth calculation. Therefore it possibly avoids, regions with missing information due to occlusions or disparity rapid changes in disparity. It is possible, that observers fixated towards the object centers, instead of object boundaries (where can be depth discontinuities). One limitation of the study might be the ground truth quality of disparity maps. The disparity maps were computed using a simple correspondence algorithm. Therefore, the noise in the disparity maps might have affected final results.

The idea that depth can be estimated from a stereo image pair based on salient features was discussed by Aziz *et. al* [Aziz 10]. They proposed biologically plausible stereo processing leading to object-based disparity computation, where object borders were used as salient features to estimate depth contrast. This idea contradicts findings of Liu *et. al* [Liu 10], however those depth-wise saliency locations restrict depth analysis and therefore reduce the computational complexity.

In 2011, Huynh-Thu *et. al* [Huynh-Thu 11] studied the differences in visual attention

when viewing of 2D and 3D videos. In this experiment, twenty-one videos with a range of disparity values and a wide variety scene changes were shown to observers both in 2D and 3D version. Fixations were recorded in a free-viewing task. The data only from the left eye was used for comparison and corresponding hitmaps were computed. Similarity of fixations in 2D and 3D were compared and some differences in fixation patterns were found. The average saccade velocity was higher for 3D videos. The average fixation frequency was always similar or higher in 2D videos. The average fixation duration was always higher in 2D videos. Furthermore, the observations reported did not indicate a wider spread of fixations in 3D videos. No strong evidence of the opposite was found either. It was found that similarity between two maps is highly content dependent. It was reported that fixation spread was highly depended on the content and video flow. In a video with a single static scene, viewers explored more areas and fixations were more widespread in 3D. However, in videos containing fast motions and rapid scene changes, the spatial extend and fixation density were similar both in 2D and 3D and often biased toward the center of the image. In some cases people were attracted by background areas in 3D but not in 2D. This was especially true when sufficient time to explore the scene was given. It was also found that high cognitive information (*i.e.* text and faces) attract human attention. In this cases fixation patterns in 2D and 3D were similar. Finally, the authors reported that, even for similar fixations in 2D and 3D, the temporal order was different.

Wang *et. al* [Wang 12] conducted binocular eye-tracking experiments by showing still synthetic stimuli on a stereoscopic display and studied a so-called 'depth-bias' in the free-viewing task. It was found that regardless of the number of objects contained in the scene, the object closest to the observer always attracts most fixations, meaning that a clear 'depth-bias' exists in the beginning of observation. The number of fixations on each object decreases as the depth increases. However, the furthest object received a couple of more fixations than several objects in front of it. As the observation time increases, the number of fixations located on the closest object becomes smaller and the distribution of fixations on all the objects in a scene becomes more uniform. These results are consistent with the result of Jansen *et. al* [Jansen 09] and indicate the existence of an additional depth prior when viewing 3D content. This location prior possibly means that depth should be integrated as an additional function.

Lang *et. al* [Lang 12b] conducted a comparative and systematic study of visual saliency in 2D and 3D scenes. Eye-tracking data was collected to create hitmaps. Those hitmaps globally represent the spatial distribution of human fixations in 2D and 3D versions. It was found that depth cues modulate visual saliency to a greater extent at farther depth ranges (spatial extent). Furthermore, humans fixate preferentially at closer depth ranges. A few interesting objects account for the majority of the fixations and this behavior is consistent across both 2D and 3D. They also found out that the relationship between depth and saliency is non-linear and is a characteristic for low and high depth-of-field scenes. Those results are consistent with those from Jansen *et. al* [Jansen 09], Huynh-Thu *et. al* [Huynh-Thu 11] and Wang *et. al* [Wang 12]. The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially, when there are multiple salient stimuli located in different depth planes.

Gautier *et. al* [Gautier 12] studied how the disparity, the center and the depth biases

influence saliency. They used the eye-tracking dataset from Jansen *et. al* [Jansen 09]. They suggested a time-dependent computational model, which predicts saliency on static pictures. The model combines low-level visual features, center and depth bias. In their study they answered 4 questions: (1) Does binocular disparity affect spatial locations of fixated areas, center and depth biases? (2) Is the predictability of state-of-the-art bottom-up saliency models affected by stereo disparity? (3) How to model the center and depth biases effects as individual features? (4) How to include these features into a time-dependent model to predict salient regions over time? Regarding question (1) they have found that (a) the presence of disparity has a time-dependent effect (first attend to closest location); (b) the central bias is observed for both 2D and 3D, but for 3D it is much lower with time; (c) disparity influences the beginning of the viewing; To answer question (2) they checked several state-of-art saliency models ([Itti 98], [Le Meur 06], [Bruce 09]) against 2D and 3D fixation data. They found that none of the models represents a mean saliency estimation in 2D significantly different than in 3D. Center and depth bias effects (question (3)) were modeled by a linear combination of low-level visual features, center and depth biases, where weights are time-dependent and learned from fixation data. To answer question (4) they compared results by combining equally different features, based on the learned weights and just a static model and came to the conclusion that a combination of learned weights over time gives the best results. Results claim that attention is influenced when binocular disparity is introduced, i.e. humans tend to look at first at closer locations (in terms of depth) and then move their gaze widespread.

Khaustova *et. al* [Khaustova 13] studied how textures with different complexity and different disparities influence attention when viewing static images. Hitmaps from fixation data were created. Basic properties were analyzed: the length of saccades and the duration of fixations. They found that the saccade length shortens over time and and also when disparity is introduced. These results are consistent with the study of Jansen *et. al* [Jansen 09] who reported reduced saccade length in 3D and overall shortened saccades over time. No proof was found that texture influences saccade length. The analysis revealed no correlation between depth levels and fixation durations. These results neither support findings of Huynh-Thu *et. al* [Huynh-Thu 11] nor Jansen *et. al* [Jansen 09] who reported that a disparity cue shortened the median fixation duration. There was no significant influence of texture complexity on fixation duration as well. Analysis of the heat maps for first 20 seconds did not indicate any significant differences in fixation patterns for different disparities or for different texture complexities. This means that depth levels did not have an obvious influence on the spread of gaze points and depth levels did not influence the selection of salient features in a scene. In general, fixations were centered in the middle and denser during the first 4 seconds, but were spread later over the entire scene. Authors came to the conclusion that object saliency plays a more important role than the depth. Moreover, areas of interest depend on the texture. This finding is supported by Vishwanath and Kowler [Vishwanath 04], who came to the conclusion that saccades tend to land closer to the object center of gravity.

Desingh *et. al* [Desingh 13] investigated the role of depth in saliency detection in the presence of (1) competing saliencies due to appearance, (2) depth blur and (3) center-bias for indoor images. Instead of using eye-tracking data they created a ground truth using a

region based method similar to [Li 13]. Eight subjects were requested to draw bounding borders around objects/regions that attracted them in the image. The ground truth value of the pixel was set to 1 if at least 2 subjects agreed that the pixel belongs to a salient region and zero otherwise. Experiments supported the hypothesis that depth influences attention in the presence of blur and center-bias. In particular, it was shown that low contrast objects at closer depth get more attention in the initial couple of seconds, which supports the hypothesis of the temporal characteristics of visual attention. It was also found that humans fixated on objects that are focused irrespectively of whether those objects have low contrast with respect to the surroundings or not. Low contrast objects placed at the center of the field of view also get more attention compared to other locations. Therefore, the notion of center bias is also applicable in large depth-of-field scenes.

As can be seen from the description above and the comparison Table 2.1 there are still many open research questions related to eye-tracking in 3D. It is clear, that the gaze is attracted not only by 2D features, such as the mean luminance, the luminance contrast and the luminance gradient, but also such features as the depth gradient and the depth contrast may influence human fixations. A significant number of researchers agree that people tend to fixate at first on closer objects and then gradually shift their gaze to objects situated further away. The majority of researchers agree, that when 3D information is involved, fixations are spread widely, especially for static scenes. There is also evidence that center bias exists not only when viewing 2D images, but also 3D. Some researchers report that humans tend to fixate on objects, and not on specific features in the scene. Also, high level cognitive information, such as texts and faces, attract human attention. There are still many unclear details concerning eye-tracking in 3D. Different researchers used different data and metrics to report their results. This potentially can mean that those results cannot be fully compared and more experiments are needed to make certain conclusions before human eye-tracking patterns tend to be different in indoor and outdoor environments. However, up to this point it is obvious that depth is involved in the human attentional process and therefore should be considered when computational attention models are designed.

The above findings can be efficiently utilized in robotics, for example by bringing attention to closer objects first in order to avoid obstacles and to start visual processing from the nearest parts of the scene. Also, the evidence that humans attend to complete objects in 3D is of a high interest for robotics, because identifying objects is a part of many robotic tasks.

It is worth mentioning that to the best of our knowledge there are no eye-tracking experiments conducted in real 3D environments, rather much of it is done in front of a monitor with 3D images. One of the first attempts to investigate attention in real environments was implemented by Pirri *et. al* [Pirri 11]. They described a novel approach to online 3D gaze estimation and their proposed approach uses a general gaze model to calculate the Point of Regard in the 3D world. They showed how this approach can be applied effectively to examine visual attention by estimating gaze in experiments that involve people performing every day tasks in natural environments.

Later, Paletta *et. al* [Paletta 13] described a vision system that is capable of mapping human attention onto a previously 3D reconstructed real-world environment. The proposed methodology enabled offline full 3D recovery of the gaze directly into an automatically

computed 3D model. This approach brings new potential into research on eye-tracking in 3D, opening new possibilities for attention studies.

However, no statistically valuable eye-tracking experiments were performed using the two described approaches. Conducting such experiments will bring significant contribution into research on visual attention, as well as open new perspectives for robotic visual systems.

2.1.3 Attention Operators for RGB-D Images

We continue with the discussion of existing models for 3D visual attention and explain their merits and limitations for robotic vision. The overview of all described methods that use 3D information to calculate saliency map is given in Table 2.2. We conclude with a brief description of methods proposed in this thesis to solve existing problems in robotic attention systems.

In general all saliency models for RGB-D images can be divided into three categories based on the way depth information is integrated into the attentional model: (1) models that use depth weighting, (2) models that extract features from depth, and (3) models based on the human eye-tracking data in 3D. We will start our discussion from the first type of models – depth-weighting models, simply because they are more straightforward than others and appeared first in literature.

However, before we go deep into details of attention models, we would like to explain how those models are typically evaluated. Probably, the best way to evaluate visual attention models is to compare the produced saliency maps against human fixation maps. There is no standardized measure to compute the similarity, however the most common metrics include: 1) Pearson linear correlation coefficient (PLCC) [Le Meur 06], 2) the area under the receiver operating characteristics curve (AUC) [Zhang 08], 3) Kullback-Leibler divergence metric (KLD) [Bruce 05]. PLCC and KLD metrics can be directly used to compare fixation maps and saliency maps, while AUC can be applied to compare fixation points to saliency maps. These comparison methods have several significant drawbacks. The first, and probably the most obvious problem, one may face is the lack of eye-tracking data. This problem is especially relevant in cases when saliency map computation involves depth information, because until the time of research there exist only a few eye-tracking datasets for 3D images (see 2.1.2). The second problem is concerned with the fact that algorithms for computational saliency developed for domestic robotics intend to operate in man-made indoor environments, while eye-tracking experiments are mostly performed on natural scenes. This becomes a problem because gaze patterns can significantly differ between natural scenes and indoor environments. The last, but not the least, an attention system for robotics does not necessarily need to mimic the human visual system, but rather concentrate on solving visual tasks such as detecting objects and obstacles. The ideal solution will be to have a system that solves detection tasks accurately and is comparable to the performance of a human.

Due to the above mentioned problems in evaluating attention models against eye-tracking data, researchers in visual attention tend to evaluate proposed saliency maps using different metrics. One of these metrics is the number of correctly detected objects in the

PAPER		[Maki 96]	[Backer 00]	[Courty 03]	[Frintrop 05b]	[Lee 05]	[Bruce 05]	[Higuli 05]	[Deville 08]	[Y.-M. 07]	[Akman 10]	[Zhang 10]	[Niu 12]	[Lang 12b]	[Wang 13]	[Desingh 13]
CC		—	●	—	—	—	—	●	●	●	—	—	●	—	—	—
O/OC		—	—	●	—	—	—	●	—	●	—	—	—	—	—	—
IC		—	—	—	—	—	—	●	●	●	—	—	—	—	—	—
EC		—	●	—	—	—	—	—	—	—	—	—	—	—	—	—
SYM	**2D**	—	●	—	—	—	—	—	—	●	—	—	—	—	—	—
	3D	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
D(3D)	**PR**	—	—	—	—	—	—	—	—	—	—	—	—	●	●	—
	DC	—	—	—	●	—	—	●	●	—	—	—	●	—	●	—
	OC	—	—	—	●	—	—	—	—	—	—	—	—	—	—	—
	AB	—	—	—	—	—	—	—	—	—	—	—	●	—	—	—
	DG	—	—	—	—	—	—	—	●	—	—	—	—	—	—	—
	B	●	●	●	—	—	●	—	—	●	●	●	●	—	●	●
	C/N	—	—	—	—	●	—	—	—	—	●	—	—	—	—	●
M/IF		●	—	—	—	—	—	—	—	●	—	●	—	—	—	—
CB		—	—	●	—	—	—	—	—	—	—	—	—	—	—	—
MULTI-LEVEL		—	—	—	●	●	—	●	●	●	●	●	—	—	—	—
COMB.	**NN**	—	●	—	—	—	●	—	—	●	—	—	—	—	—	—
	$\oplus$	—	—	●	●	●	—	●	●	●	●	●	●	●	●	—
	$\otimes$	—	—	—	—	—	—	—	—	●	—	●	●	●	●	—
	SVM	—	—	—	—	—	—	—	—	—	—	—	—	—	—	●[5]
REMARKS		[1]	—	—	—	—	—	—	—	[2]	—	[3]	[4]	—	—	[4]
COMB./W		—	—	—	—	—	—	—	—	—	—	T.2.3	—	T.2.3	T.2.3	T.2.3
DB		T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3
EVALUATION	**Q**	●	●	●	●	●	●	—	—	—	●	●	—	—	—	—
	PR	—	—	—	—	—	—	—	—	—	—	—	●	—	—	—
	OD	—	—	—	—	—	—	—	●	●	—	—	—	—	—	—
	ET	—	—	—	—	—	—	●	—	—	—	—	—	●	●	●[6]
	SM	—	—	—	T.2.3	—	—	T.2.3	—	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3	T.2.3

O|CC/IC/OC/DC|DG	**O**rientation|**C**olor/**I**ntensity/**O**rientation/**D**epth **C**ontrast|**D**epth **G**radient
SYM/EC/M/IF/CB/B/AB/PR|C/N	**SYM**metry/**EC**centricity/**M**otion and Dynamics/**I**mage **F**low/**C**entral **B**ias/Depth **B**ias,Weighting,Value/**AB**ruptness/**PR**iors|**C**urvature/**N**ormals
NN/⊕/⊗/SVM	**N**eural **N**etwork/accross-scale addition/multiplication/**S**upport **V**ector **M**achine
Q/PR/OD/ET/SM	**Q**ualitive/**P**recision–**R**ecall curve/**O**bject–**D**etection/**S**aliency **M**odels

[1] target masks are produced from features

[2] ICA to integrate features

[3] correlation between channels is taken into account

[4] region-based saliency

[5] there are multiple features in the feature vector, see text for details

[6] region-based method similar to [Li 13] instead of eye-tracking

Table 2.2: The table shows comparison of different attention algorithms developed to be used with RGB-D data.

PAPER	DATABASE	[Itti 98]	[Zhai 06]	[Harel 06]	[Hou 07]	[Achanta 09a]	[Bruce 09]	[Seo 09]	[Achanta 10]	[Goferman 10]	[Cheng 11]	[Lang 12a]
[Maki 96]	STEREO VIDEO SEQUENCES	—	—	—	—	—	—	—	—	—	—	—
[Backer 00]	VIRTUAL RENDERED SEQUENCES	—	—	—	—	—	—	—	—	—	—	—
[Courty 03]	VIRTUAL RENDERED SEQUENCES	—	—	—	—	—	—	—	—	—	—	—
[Frintrop 05b]	LASER SCAN IMAGES	**E**	—	—	—	—	—	—	—	—	—	—
[Lee 05]	MESH MODELS	—	—	—	—	—	—	—	—	—	—	—
[Bruce 05]	STEREO PAIRS	—	—	—	—	—	—	—	—	—	—	—
[Hügli 05]	3D IMAGES	**E**	—	—	—	—	—	—	—	—	—	—
[Deville 08]	STEREOSCOPIC VIDEO SEQUENCES	—	—	—	—	—	—	—	—	—	—	—
[Y.-M. 07]	VIDEO SEQUENCES	**E**	—	—	—	—	—	—	—	—	—	—
[Akman 10]	POINT CLOUDS FROM TOF CAMERA	—	—	—	**E**	**E**	—	—	—	—	—	—
[Zhang 10]	MULTI-VIEW VIDEO SEQUENCES	**C/E**	—	—	—	—	—	—	—	—	—	—
[Niu 12]	STEREOSCOPIC IMAGES	—	—	**E**	**E**	**E**	—	—	**E**	**E**	**E**	—
[Lang 12b]	NUS-3DSALIENCY	**C/E**	—	**C/E**	**C/E**	**C/E**	**C/E**	**C/E**	—	—	—	**C/E**
[Wang 13]	MIDDLEBURY; IVC 3D IMAGE	**C/E**	—	—	**C/E**	—	**C/E**	—	—	—	—	—
[Desingh 13]	UW DB, BERKLEY DB	—	**C/E**	—	**C/E**	—	—	—	—	—	**C/E**	—

C – **C**ombination **E** – **E**valuation

Table 2.3: The table shows how different state-of-the-art attention operators, developed for RGB-D data were evaluated and/or combinated with state-of-the-art 2D saliency algorithms.

scene. Given a saliency map, a set of so-called fixation points or regions is extracted. These fixation points or regions are then compared to the ground truth objects and the conclusion about the model performance can be made. This approach is especially popular in robotics, due to its simplicity and practical motivation. Moreover, this approach allows to evaluate the robustness of the attention system under different environments and conditions.

In the field of computer vision, saliency maps are usually evaluated by means of a *PR*-curve (Precision-Recall curve). This approach is used when saliency maps are designed to detect the most attractive object in the scene. Given a saliency map and the ground truth segmentation of the object of interest, the saliency map is thresholded and precision and recall metrics of the resultant segmentation are calculated. By varying the threshold value, a *PR*-curve can be built. One of the drawbacks of this evaluation method is that it can be efficiently used only when there is only one attractive object in the scene. This approach is not very popular in robotics, because in typical robotics applications the number of objects of interest can be unlimited.

Depth-Weighting Saliency Maps

One of the first attention models that used a depth-weighting scheme was proposed in 1996 by Maki *et. al* [Maki 96]. The architecture is based on a parallel detection of pre-attentive

cues (image flow, stereo disparity, and motion detection), followed by a cue integration mechanism based on the selection criteria using nearness and motion. Two masks based on the pursuit and saccade modes are computed then. Depth is applied as a priority criterion to select the mask. The hypothesis assumed by Maki *et. al* [Maki 96] is that closer target has higher priority. As explained by the authors, this hypothesis is valid in scenarios where the observer has to avoid obstacles. Indeed, they demonstrated the application of the model for a scenario with moving or stationary stereo-camera. The camera selectively focuses (masks) on different moving objects and holds focus on them for several frames. It was shown that the model kept fixating on the moving object closest to the camera. For the evaluation they did not provide any database with ground truth, but merely demonstrated results on some video sequences with moving people.

Backer *et. al* [Backer 00] introduced the Neural Active VISion system (NAVIS). The system contained a bottom-up component, which took into consideration such pre-attentive cues as eccentricity, color contrast and symmetry. It also integrated depth and time to get a 3D saliency representation for appropriate behavior in a three-dimensional dynamic world. In the paper authors focused mostly on the mechanism of attentional control itself and did not give any evidence that the system can be used in the real world environment. To show how the proposed system works they applied it on a series of rendered sequences consisting of a number of disjoint objects.

In 2003, Courty *et. al* [Courty 03] showed how saliency maps can be used for image rendering purposes. The hypothesis that was explored is that close objects should be more salient for humans. The saliency map was based on a set of response maps from Gabor filters and a depth image. A Gaussian image was subtracted from the saliency map to simulate the center bias. A list of fixations was extracted from the saliency map and then used as input to the animation engine. They tested the proposed algorithm for saliency computation on a virtual character wandering along the street in a virtual city.

Jang *et. al* [Y.-M. 07] based their attention model on static and dynamic saliency maps. To obtain the static map intensity, orientation, color and symmetry feature maps are calculated. The symmetry feature map is obtained from orientation maps. All feature maps are then integrated into the static saliency map by an independent component analysis (ICA) algorithm based on entropy maximization. The dynamic saliency map is based on the dynamic analysis of successive static saliency maps. The authors proposed to use depth information in a decaying exponential function meaning that closer objects attract more attention than those situated far away. Evaluation was carried qualitatively to show that the proposed attention model performs better than a classical model of Itti *et. al* [Itti 98].

Zhang *et. al* [Zhang 10] proposed a bottom-up visual attention model for stereoscopic content. They extended a hierarchical model by Itti *et. al* [Itti 98] for stereoscopic vision by using the depth map and motion as an additional cue. Optical flow is utilized to estimate motion of image objects between consecutive frames. The motion map is decomposed with a dyadic pyramid into nine levels and center-surround difference is adopted to separate the attentive object motion from background motion. Near objects are considered more important than far objects. All maps are then added together, while the correlation between channels is subtracted, and then the final saliency map is multiplied with the depth saliency. As evaluation, they performed attention detection experiments on multi-view video sequences.

However, no quantitative evaluation was performed.

We can raise two main criticisms concerning the above described works by Maki *et. al* [Maki 96], Backer *et. al* [Backer 00], Courty *et. al* [Courty 03], Jang *et. al* [Y.-M. 07], Zhang *et. al* [Zhang 10]. Firstly, the proposed approaches were designed either for rendering purposes or for detecting closer objects as more salient. These systems can be efficiently used in the task of autonomous driving or obstacle avoidance. However, the hypothesis that closer objects attract more attention may not necessarily hold in a scenario of viewing complex scenes where the closest object may not be the only or main object of interest. These approaches internally assume that objects in the scene are disjoint and are not situated in clutter, while robots have to face cluttered environments with multiple occluded objects. In the case of a clutter, attention in these systems will be driven to the closest regions of the clutter and those areas do not yield any information regarding object boundaries and hence cannot localize objects. Therefore, these approaches cannot be directly applied to robotics applications for detecting objects in cluttered environments. Secondly, while the hypothesis that closer objects are more salient has been proven to be biologically plausible, it still cannot serve as the only one depth feature when building attention systems for a robot. A more complicated scheme must be involved to extract attention features from the depth channel in order to prioritize objects relevant to the task at hand.

Saliency Maps from Depth-Extracted Features

The second group of attentional models extracts features directly from depth and uses them in the process of saliency computation. In 2005, Frintrop *et. al* [Frintrop 05b] proposed to extend the hierarchical model described by Itti *et. al* [Itti 98] to be used with depth and reflectance images from laser scans. In this model image pyramids with five levels each, were created from depth and reflection. Intensity and orientation feature maps were calculated using on-center-off-surround and off-center-on-surround operators. Those feature maps were fused together to obtain a master saliency map. The qualitative evaluation showed that the proposed model solves the task of detecting salient objects better than the model of Itti *et. al* [Itti 98].

Lee *et. al* [Lee 05] applied the idea of Itti *et. al* [Itti 98] to 3D meshes. The mean curvature at each vertex is used as the feature to measure saliency. The scale space is defined as the size of the neighborhood of the vertex in the mesh. The saliency model was applied to the task of mesh simplification and salient viewpoint selection. For qualitative evaluation, they used five mesh models.

Hügli *et. al* [Hügli 05] (also Jost *et. al* [Jost 05] and Ouerhani *et. al* [Ouerhani 00]) extended the model proposed by Itti *et. al* [Itti 98] with depth features. They hypothesized that the depth contrast contributes to visual attention similarly to intensity contrast, therefore a model with depth features predicts human fixations better than a model without. In the proposed model a number of conspicuity maps are built from intensity, color and depth low-level features using a multi-resolution center-surround mechanism. A linear combination of conspicuity maps is used to produce an overall saliency map for the image. To evaluate the proposed model they recorded eye movements for 12 stereo images containing natural indoor and outdoor scenes. They calculated similarity between human fixations

and produced saliency maps. The usefulness of integrating depth information in a computational model was proven by increase in the similarity with human fixations up to 10% when the depth feature was involved.

In a way similar to Hügli *et. al* [Hügli 05], Deville *et. al* [Deville 08] used intensity, color, depth and depth gradient features to create a saliency map using a multi-resolution center-surround mechanism. However, coefficients for linear combination of conspicuity maps were based on the size of regions of conspicuity, i.e. as a function of the mean size of conspicuous regions. They carried out experiments in two different situations: (1) indoor scenes with obstacles (like furniture and people), and (2) outdoor street scenes (with outdoor object and pedestrians). The ground truth for objects that are desired to be detected was determined manually. As a quantitative measure they evaluated the success rate of detecting ground truth obstacles.

Akman *et. al* [Akman 10] proposed a novel saliency detection mechanism based on the combination of geometric and spatial information from the 3D point cloud data. They utilized local surface properties to detect non-trivial regions in a scene. At first, five spatial scales of the input point cloud are generated using dyadic Gaussian pyramids with bilinear interpolation. Non-trivial regions or spatial irregularities are detected at each point on every scale by means of a residual, the minimum eigenvalue of a covariance matrix for plane fitting at each point. Calculated residual values at each scale are averaged to create a saliency map. This saliency map is linearly combined with a second saliency map based on the exponential decay of depth values to obtain a master saliency map. They visually compared the proposed saliency map on 3D point clouds with saliency maps by Hou *et. al* [Hou 07] and Achanta *et. al* [Achanta 09a].

The above described methods by Frintrop *et. al* [Frintrop 05b], Lee *et. al* [Lee 05], Hügli *et. al* [Hügli 05], Deville *et. al* [Deville 08] and Akman *et. al* [Akman 10] propagated the idea of detecting saliency by finding irregularities, such as depth or curvature contrasts, in the scene. It means that regions that have different depth characteristics compared to their environment will become salient. However, this approach only works if objects are separated from each other in space or some objects have significantly different shape characteristics compared to their neighborhood. In cases of severe clutter with multiple similar objects, the saliency will yet again be defined only by object color properties.

Bruce and Tsotsos *et. al* [Bruce 05] described the problem of binocular rivalry appearing in stereo-based vision and the expansion of attention to 3D. They described the difficulty in biologically plausible translation of 2D visual attention models to the stereo-based vision. In detail, they demonstrated how hierarchical models can be used to extract features to create saliency maps. Those saliency maps are then used to predict fixations. Authors pointed out that it is not completely biologically plausible to extract features independently. Moreover, it is significantly more complicated to prove independent feature extraction to model stereo-based attention. It is mentioned that stereo-based attention models should consider conflicts between the two eyes that often result in occlusions or higher disparity values. Therefore, two eyes and it corresponding fixations should be coupled. Moreover, 2D saliency map eliminates the correspondence between the eyes. Following above given arguments, they suggested a stereo-based attention approach following 2D visual attention model that uses an architecture based on the pyramid processing. Inter-

pretive neuronal units were added to the pyramid to achieving stereo-vision. The study showed several synthetic images with structural objects, but did not suggest any results in more complex scenarios. Unfortunately, results are not compared with ground-truth data from eye-tracking experiments.

In 2012, Niu *et. al* [Niu 12] described an algorithm for stereo saliency using disparity information. In the proposed algorithm, an RGB-D image is first segmented into uniform regions based on color and disparity. The saliency value of each region is then calculated. The authors proposed two different types of saliency: (1) saliency based on global disparity contrast and (2) saliency based on domain knowledge. Global disparity contrast-based saliency was extended from the color contrast-based saliency proposed by Cheng *et. al* [Cheng 11]. In the algorithm, disparity difference between regions is computed instead of color-based difference; abruptness of disparity change, i.e. maximal disparity change along the path between two pixels, is considered as an additional feature. The domain knowledge saliency map consists of two maps: (a) saliency map where objects with small disparity magnitudes (comfort zone) tend to be more salient and (b) saliency map where objects popping out from the screen tend to be more salient. Then those two maps are linearly combined and modulated using the local disparity contrast in each row to obtain the domain knowledge saliency map. Final stereo saliency is obtained by multiplying Domain knowledge saliency map and global disparity contrast-based saliency map. The authors evaluated their algorithm on a collection of 1000 stereoscopic images, where the most salient object in each image was labeled manually by three users. They showed that the proposed stereo saliency map performs better in terms of PR compared to existing state-of-the-art attention models ([Harel 06, Hou 07, Achanta 09a, Achanta 10, Goferman 10, Cheng 11]). The approach was designed for the task of automatic object segmentation by taking binarized saliency map as input and can be efficiently used when there is only one object of interest in the image. This limits the application of the algorithm in robotics, where several objects can be objects of interest simultaneously.

Saliency Maps based on Eye-Tracking

The last group of attentional models uses eye-tracking data to learn depth-priors for saliency computation. Lang *et. al* [Lang 12b], based on eye-tracking experiments (see 2.1.2), proposed to model the relationship between depth and saliency by approximating the joint density with a Mixture of Gaussians. The model parameters were obtained from a training dataset and the Expectation-Maximization (EM) algorithm was applied for fitting Gaussian mixtures. They extended seven existing methods ([Itti 98, Harel 06, Hou 07, Achanta 09a, Bruce 05, Seo 09, Lang 12a]) to include the learned depth prior using addition and multiplication operators. The evaluation was carried on the collected fixation dataset in terms of ROC and AUC metrics, as well as the correlation coefficient (CC) between the fixation map and the predicted saliency map. The evaluation showed that models with predicted depth priors perform significantly better than those without such depth priors. It was found that multiplicative modulating effect explains the influence of depth on saliency better than a linear weighted summation model. The predicted depth saliency maps are similar in their spatial distribution between 2D and 3D versions when there is one conspicuous area or

object clearly standing out from others, but when the scenes include multiple objects or no conspicuous objects, there is a noticeable difference between the predicted depth saliency maps in 2D and 3D cases. The authors came to the conclusion that extending the saliency models to include the proposed depth priors consistently improves the performance of current state-of-the-art saliency models.

Wang *et. al* [Wang 13] proposed a depth-based saliency model of 3D visual attention. They applied Bayes theory on the result of their eye-tracking experiments using synthetic stimuli *et. al* [Wang 12] (see Sec. 2.1.2) to model the correlation between depth features and the levels of depth saliency. Their framework can combine the obtained depth saliency map with existing 2D visual attention models. They used only depth contrast as a feature and only one scale because they claimed there is no proof that there exists multi-scale depth perception. Three bottom-up visual attention models ([Itti 98, Hou 07, Bruce 05]) were used as 2D saliency predictors and then combined with depth saliency map. The final saliency map was obtained by addition or multiplication of 2D saliency map and depth saliency map. To evaluate the proposed model the authors created a new eye-tracking database containing 18 stereoscopic natural content images. They evaluated each 2D model, the proposed model, combinations of 2D models with addition and multiplication of the proposed model, and 2D models with multiplication by depth, in terms of correlation between ground-truth eye-tracking data and resulting saliency maps. Experiments supported the suggestion to consider depth information as an individual cue to model visual attention. Combining the depth saliency map with the 2D models showed good results for prediction of salient areas.

Desingh *et. al* [Desingh 13] developed a 3D-based saliency model that integrates depth and geometric features of object surfaces in indoor scenes. Their approach consists of creating two types of saliency maps: (1) color-based and (2) depth-based. Those two saliency maps are later fused together into a master saliency map. To create the depth-based saliency map they segment the input point cloud using a region growing technique with curvature and smoothness of the surface as features. The saliency of each region is computed using contrast score based histograms of surface normals, taking into consideration the number of points in the region. For the color-based saliency map, they used several state-of-the-art algorithms ([Zhai 06, Hou 07, Achanta 09a, Cheng 11]). Fusion of color-based and depth-based saliencies was done through non-linear regression using a Support Vector Machine (SVM). Given the saliency values in color-based and depth-based maps, the SVM predicted the saliency value in the final map. They also used additional features of regions, such as color histogram, contour compactness, dimensionality, perspective score, discontinuity, size and location, verticality. The ground truth for learning was obtained by a region-based method, similar to [Li 13] instead of eye-tracking. For evaluation images from the University of Washington RGB-D dataset, the Berkeley 3D object dataset, and their own dataset were used. It was shown that the proposed saliency model outperforms existing attention models in terms of Receiver-Operator-Characteristics.

The above described models of Lang *et. al* [Lang 12b], Wang *et. al* [Wang 13] and Desingh *et. al* [Desingh 13] were based on eye-tracking data. Moreover, the depth-based saliency map was always considered as a complimentary feature to existing color-based saliency models. While the human visual system is significantly superior to any existing

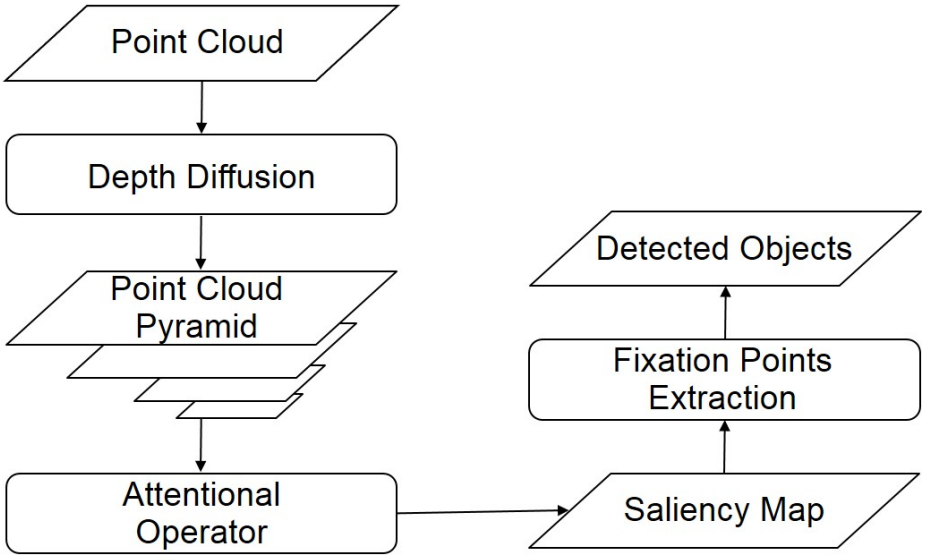


Figure 2.1: A general scheme to create saliency maps from an input point cloud. At first point cloud is diffused to obtain missing depth data. Then a Gaussian pyramid of the point cloud is created to enable computations on different scales. Attentional operator is applied on the pyramid to create a saliency map. Finally fixations points that represent object candidates are extracted from a saliency map.

artificial visual system, still these systems are intended to solve different tasks. This means that what can be salient for humans does not necessarily need to be salient for a robot. Therefore, attentional operators based on eye-tracking data may be good in predicting human fixations, but do not always serve as good predictors for objects, as will be shown in our evaluation Section.

The goal of this thesis is to create attentional operators for a robotic system that are capable of reliably detecting objects in highly cluttered indoor environments. We show that the proposed approach works better compared to existing state-of-the-art algorithms in terms of the quality and precision of object detection. In the next Section we describe in detail the proposed attentional models and later continue with the evaluation showing the benefits of the proposed algorithms.

2.2 Object Detection using Saliency Maps

A general scheme to detect objects using saliency maps is presented in Fig. 2.1.

As input, each saliency operator takes an organized RGB-D point cloud $\mathbf{P}$. Each point $\mathbf{r}_i \in \mathbf{P}$ is represented using a set of coordinates

$$\mathbf{r}_i = (r_i, c_i, \mathbf{p}_i, \mathbf{n}_i) \tag{2.1}$$

where (r_i, c_i) are row-column coordinates in the image, $\mathbf{p}_i = (x, y, z)$ the 3D point coordinates, and $\mathbf{n}_i$ the unit normal. For each point $\mathbf{p}_i$, the unit normal $\mathbf{n}_i = (n_{ix}, n_{iy}, n_{iz})$ is estimated as the normal of a plane tangent to the neighboring surface [Rusu 09].

We create a depth image D from the point cloud $\mathbf{P}$, so that $\forall \mathbf{r}_i \in \mathbf{P}$:

$$D\left(r_i, c_i\right) = z \tag{2.2}$$

Because of missing depth values in the depth image acquired by Kinect, due to shadows, black regions, shiny surfaces, transparent materials and many others, we first perform Markov Random Field (MRF) guided anisotropic depth diffusion [Varadarajan 12].

After that a Gaussian pyramid is created to enable calculation of multi-level saliency map. This process is described in Section 2.4. We propose a set of different saliency operators for RGB-D data, which are described in Section 2.3. Object detection is implemented by means of extracting fixation points (they are also called attention points), from a saliency map. Extraction strategies that are proposed to be used in this thesis are described in Section 2.5. Finally, in Section 2.6 we evaluate the proposed saliency maps in combination with different extraction strategies against existing state-of-the-art attentional operators.

2.3 Saliency Maps for RGB-D

In this Section, we are going to discuss our proposed saliency operators:

- *Point-based Height Saliency Map (PH)* (2.3.1)
- *Surface-based Height Saliency Map (SH)* (2.3.2)
- *Relative Surface Orientation Saliency Map (RSO)* (2.3.3)
- *3D Symmetry-based Saliency Map (SYM3D)* (2.3.4)

2.3.1 Point-based Height Saliency Map

In this Section, we describe the point-based height saliency map S_{PH}. Height above the supporting plane is considered to be one of the attention cues that grasps humans visual attention. Moreover, object height can be a useful feature in such tasks as picking up objects, since the simplest way to start grasping is to pick up at first objects standing out from the clutter. Tall objects can be detected by attention points derived from the point height saliency map (Fig. 2.10).

Given a point cloud $\mathbf{P}$, we want to detect the height of each point $\mathbf{r}_i \in \mathbf{P}$. To calculate height, we need to determine a reference, i.e. the supporting plane χ on which objects rest (e.g. a table). From the task context, such as grasping objects from a table we can assume that such supporting plane exists. The plane χ with the equation in the Hesse normal form $\mathbf{n}_\chi \mathbf{p} - \rho = 0$ is determined using RANSAC [Fischler 81] algorithm, where $\mathbf{n}_\chi$ is the unit plane normal. Only points whose projections on the table are situated inside the table convex hull are used to create a saliency map.

For every point $\mathbf{r}_i$ its unsigned Euclidean distance $d_\chi\left(\mathbf{r}_i\right)$ to the supporting plane χ is calculated according to:

$$d_\chi\left(\mathbf{r}_i\right) = \left|\mathbf{n}_\chi \mathbf{p}_i + \rho\right| \tag{2.3}$$

The point-based height saliency value of the point $\mathbf{r}_i$ is equal to:

$$S_{PH}\left(r_i, c_i\right) = \kappa {d_\chi}^2\left(\mathbf{r}_i\right) \tag{2.4}$$

A squared distance from the plane is used to obtain more pronounced saliency maps, and κ is chosen such that $S_{PH}\left(r_{i^*}, c_{i^*}\right) = 1$, where $i^* = \underset{i:\mathbf{r}_i \in \mathbf{P}}{\operatorname{argmax}}\, d_\chi\left(\mathbf{r}_i\right)$.

2.3.2 Surface-based Height Saliency Map

In the Section, 2.3.1 we have described the point-based height saliency map. In this Section we are going to describe the surface-based height saliency map S_{SH}. While the point-based height saliency map assigns a saliency value to each point independently, surface-based saliency map assigns saliency values based on the highest point of the local neighborhood around the point (Fig. 2.10). The motivation for this type of saliency map comes from the idea to assign the saliency value to a given point based on the saliency value of the highest point above the given one. This allows to detect "short" and "tall" objects.

To calculate height we again need to determine a reference, i.e. the supporting plane. The support plane χ with the equation in the Hesse normal form $\mathbf{n}_\chi \mathbf{p} - \rho = 0$ is determined using RANSAC [Fischler 81] algorithm, where $\mathbf{n}_\chi$ is a unit plane normal. Only points whose projections on the table are situated inside the table convex hull are used to create a saliency map.

For every point $\mathbf{r}_i$ its unsigned Euclidean distance $d_\chi\left(\mathbf{r}_i\right)$ to the supporting plane χ is calculated according to:

$$d_\chi\left(\mathbf{r}_i\right) = \left|\mathbf{n}_\chi \mathbf{p}_i + \rho\right| \tag{2.5}$$

We project each point in the point cloud $\mathbf{r}_i \in \mathbf{P}$ on the support plane χ to obtain the projected version of the point $\mathbf{r}'_i$. For every point $\mathbf{r}'_i$ we define its neighborhood $N(\mathbf{r}'_i)$ as points in a 5 mm radius around the point $\mathbf{r}'_i$. The surface-based height saliency value $SH(r_i, c_i)$ of the point $\mathbf{r}_i$ is defined as the maximum point-plane distance within the neighborhood (to the support plane χ):

$$SH(r_i, c_i) = \max_{j:\mathbf{r}'_j \in N(\mathbf{r'}_i)} \left|d_\chi\left(\mathbf{r}_j\right)\right| \tag{2.6}$$

The surface-based saliency map S_{SH} is linearly normalized to be in the range $[0, 1]$.

2.3.3 Relative Surface Orientation Saliency Map

Objects with surfaces parallel to the supporting plane often represent good object candidates for grasping, especially in the clutter. In this Section we describe the attentional operator S_{RSO} that aims to identify top-surfaces based on relative surface orientation. Fig. 2.10 shows the example of the relative surface orientation saliency map.

To calculate relative surface orientation we determine the support plane χ with the equation in the Hesse normal form $\mathbf{n}_\chi \mathbf{p} - \rho = 0$ using RANSAC [Fischler 81] algorithm, where $\mathbf{n}_\chi$ is a unit plane normal.

The saliency value of the point $\mathbf{r}_i$ is defined as the cosine between unit point normal $\mathbf{n}_i$ and unit plane normal $\mathbf{n}_\chi$:

$$S_{RSO}\left(r_i, c_i\right) = \left|\mathbf{n}_i \cdot \mathbf{n}_\chi\right| \tag{2.7}$$

The relative surface orientation saliency map S_{RSO} is linearly normalized to be in the range $[0, 1]$.

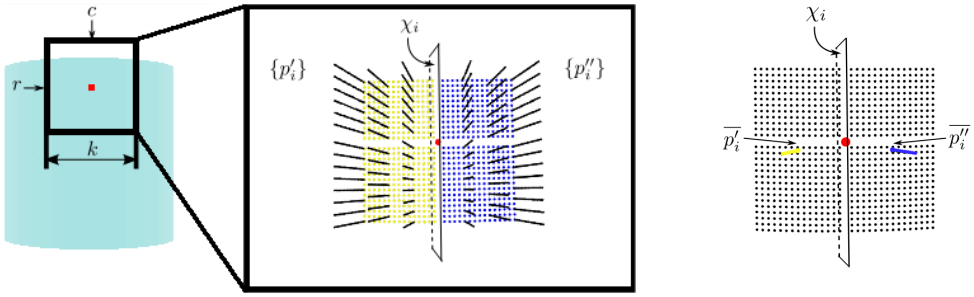
2.3.4 3D Symmetry-based Saliency Map

Symmetry is one of the characteristics of human-made and natural objects, and thus can be seen as an objectness measure. Kootstra *et al.* [Kootstra 08] showed that human eye fixations can be predicted well by symmetry. Many symmetry operators exist [Reisfeld 95, Heidemann 04, Loy 03, Chertok 10, Berner 11], which can be divided into two major groups: operators working on 2D data, and operators working on 3D data. Among the 2D operators one well-known operator was developed by Reisfeld *et al.* [Reisfeld 95] which detects context-free generalized symmetry based on magnitude orientations of image gradients. This symmetry operator was extended by Heidemann [Heidemann 04] to a local color symmetry operator. Loy and Zelinsky [Loy 03] proposed to detect local radial symmetries in an image using a special transform. Kootstra *et al.* [Kootstra 08] proposed a 2D symmetry saliency operator based on the symmetry operator by Reisfeld *et al.* [Reisfeld 95]. This saliency operator is able to detect symmetrical regions at different scales. The basic idea is that symmetries are computed over multiple scales and then summed in an across-scale addition manner to obtain a master saliency map. Kootstra *et al.* [Kootstra 08] showed that this approach works better than the classical contrast-based saliency model by Itti *et al.* [Itti 98] for scenes that contain symmetrical objects. Mitra *et al.* [Mitra 12] gave an extensive overview of the existing methods to detect different types of symmetries in 3D geometries. Methods based on search in oriented histograms [Sun 97], spectral analysis [Chertok 10], feature-graph matching [Berner 11] and many others were explored. The majority of 3D algorithms work on a complete 3D model of an object to detect symmetries. This property limits their use as bottom-up attentional operators working on RGB-D images, i.e. partial views of objects.

In this thesis, a new 3D symmetry-based saliency operator, calculating a measure of context-free local symmetry from a 3D point cloud, is proposed. The proposed symmetry operator is used to predict fixation points for further attention-driven segmentation or detailed exploration of the scene.

Our method for calculating saliency map S_{SYM3D} is based on local symmetries in 3D. The algorithm detects reflective symmetries using principal axes of Extended Gaussian Images built from normals of surface patches. The 3D reflective symmetry can be calculated from a depth image D (Fig. 2.2). $\Phi(\mathbf{r}_i)$ defines the symmetry kernel as a square patch with side length k, centered around $D\left(r_i, c_i\right)$ (Fig. 2.2). The amount of 3D symmetry at the given location $D\left(r_i, c_i\right)$ is estimated on the subset of points $\{\mathbf{r}'_i\} = \Phi(\mathbf{r}_i) \bigcap P$.

Sun *et al.* [Sun 97] proposed to use an Extended Gaussian Image built from point normals to detect symmetries of a model. Minovic *et al.* [Minovic 93] proved that planes of reflective symmetries are perpendicular to the directions of the principal axes. Thus, to



(a) (b)

Figure 2.2: The depth image of a cylinder (artificial data) is shown in Figure 2.2(a), with the point $\mathbf{r}_i$ for which the symmetry is calculated (highlighted in red), and the kernel $\Phi(\mathbf{r}_i)$ shown as a black square. The subset of points $\{\mathbf{r}'_i\} = \Phi(\mathbf{r}_i) \bigcap P$ is shown in 3D on the right side. Subsets $\left\{\mathbf{r}'^1_{ij}\right\}$ and $\left\{\mathbf{r}'^2_{ij}\right\}$ are shown in yellow and blue respectively, with the reflective plane χ_j between the two point subsets. Normals $\{\mathbf{n}_i\}$ are shown as black lines. In Figure 2.2(b) examples of $\overline{\mathbf{p}}'^1_{ij}, \overline{\mathbf{n}}'^1_{ij}$ and $\overline{\mathbf{p}}'^2_{ij}, \overline{\mathbf{n}}'^2_{ij}$ are shown in yellow and blue respectively.

detect planes of reflective symmetries from the patch we build an Extended Gaussian Image from the patch's point normals, and calculate the principal axes $\gamma = \{\gamma_1, \gamma_2.\gamma_3\}$ of the Extended Gaussian Image using Principal Component Analysis (PCA). The corresponding symmetry reflective planes χ_j $(j = 1, 2, 3)$ are defined as planes going through the point $\mathbf{r}_i$ with the plane normal equal to the corresponding principal axis γ_j.

For a given reflective plane χ_j the point set $\{\mathbf{r}'_i\}$ is divided into two subsets $\left\{\mathbf{r}'^1_{ij}\right\}$ and $\left\{\mathbf{r}'^2_{ij}\right\}$, so that $\forall \mathbf{r}' \in \{\mathbf{r}'_i\}$:

$$\mathbf{r}' \in \begin{cases} \left\{\mathbf{r}'^1_{ij}\right\} & \text{if } d_{\chi_j}(\mathbf{r}') > 0 \\ \left\{\mathbf{r}'^2_{ij}\right\} & \text{if } d_{\chi_j}(\mathbf{r}') < 0 \end{cases} \tag{2.8}$$

where $d_{\chi_j}(\mathbf{r})$ is the signed Euclidean distance from point $\mathbf{r}$ to the plane χ_j (Fig. 2.2).

The amount of 3D reflective symmetry Ω_j relative to a given plane χ_j for a given point $\mathbf{r}_i$ is defined as:

$$\Omega_j(\mathbf{r}_i) = \exp(- \triangle D_{ij}) \cdot \exp(- \triangle d_{ij}) \cdot \omega_1 \cdot \omega_2 \tag{2.9}$$

The multiplication of all four components reflects the fact, that we are searching for patches that are symmetrical in all four aspects (see below).

$\triangle D_{ij}$ represents the difference in depth values between mean points $\overline{\mathbf{r}}'^1_{ij}$ and $\overline{\mathbf{r}}'^2_{ij}$:

$$\overline{\mathbf{p}}'^m_{ij} = \frac{1}{N'^m_{ij}} \sum_{\mathbf{r}_l \in \left\{\mathbf{r}'^m_{ij}\right\}} \mathbf{p}_l \tag{2.10}$$

$$\overline{\mathbf{n}}'^m_{ij} = \frac{1}{N'^m_{ij}} \sum_{\mathbf{r}_l \in \left\{\mathbf{r}'^m_{ij}\right\}} \mathbf{n}_l \tag{2.11}$$

where $m = 1, 2$; N'^m_{ij} is the number of points in the subset $\left\{\mathbf{r}'^m_{ij}\right\}$.

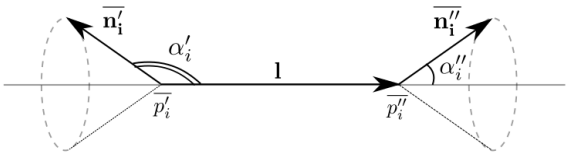


Figure 2.3: Visual illustration for the calculation of angles α_i' and α_i''. $\mathbf{l}$ is the line connecting the two mean points $\overline{\mathbf{p}}_{ij}'^1$ and $\overline{\mathbf{p}}_{ij}'^2$. α' is the angle between mean normal $\overline{\mathbf{n}}_{ij}'^1$ and $\mathbf{l}$, and α'' is the angle between mean normal $\overline{\mathbf{n}}_{ij}''^w$ and $\mathbf{l}$.

$$\triangle D_{ij} = |\overline{z}_{ij}'^1 - \overline{z}_{ij}'^2| \tag{2.12}$$

where $\overline{z}_{ij}'^1$ and $\overline{z}_{ij}'^2$ are mean depth values of the respective subsets. $\triangle D_{ij}$ reflects the fact, that we are only interested in symmetries, that are facing our view point.

$\triangle d_{ij}$ represents the difference in distances from mean points $\overline{\mathbf{r}}_{ij}'^1$ and $\overline{\mathbf{r}}_{ij}'^2$ to the reflective plane χ_j:

$$\triangle d_{ij} = || \, d_{\chi_j}\left(\overline{\mathbf{r}}_{ij}'^1\right)| - | \, d_{\chi_j}\left(\overline{\mathbf{r}}_{ij}'^2\right)|| \tag{2.13}$$

where $| \, d_{\chi_j}(\mathbf{r})|$ is the unsigned Euclidean distance from the point $\mathbf{r}$ to the plane χ_j. $\triangle d_{ij}$ reflects the fact that we are not only searching for patches with symmetrical orientations, but also for patches that can be divided into two subpatches, which are equally sized and symmetrically positioned in 3D space.

ω_1 is a coefficient measuring the co-planarity between the line $\mathbf{l}$ connecting $\overline{\mathbf{p}}_{ij}'^1$ and $\overline{\mathbf{p}}_{ij}'^2$ and the two mean normals $\overline{\mathbf{n}}_{ij}'^1$ and $\overline{\mathbf{n}}_{ij}'^2$ (Fig. 2.3):

$$\omega_1 = |[\overline{\mathbf{n}}_{ij}'^1 \times \overline{\mathbf{n}}_{ij}'^2] \times \mathbf{l}| \tag{2.14}$$

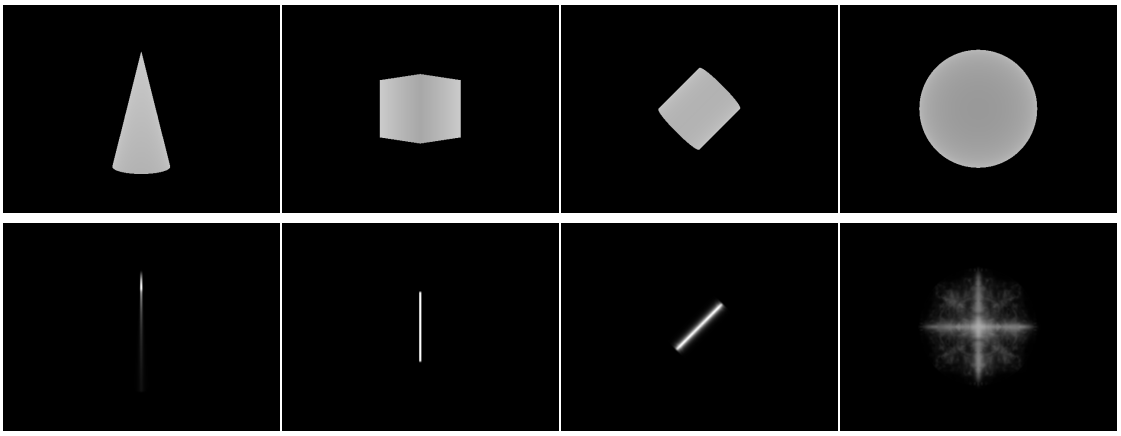
$$\mathbf{l} = \frac{\overline{\mathbf{p}}_{ij}'^1 - \overline{\mathbf{p}}_{ij}'^2}{\|\overline{\mathbf{p}}_{ij}'^1 - \overline{\mathbf{p}}_{ij}'^2\|} \tag{2.15}$$

ω_2 shows the similarity between mean normal directions based on the symmetry operator from Reisfeld *et al.* [Reisfeld 95] and is calculated as following (Fig. 2.3):

$$\omega_2 = (1 - \cos(\alpha' + \alpha'')) \cdot (1 - \cos(\alpha' - \alpha'')) \tag{2.16}$$

where α' is the angle between mean normal $\overline{\mathbf{n}}_{ij}'^1$ and $\mathbf{l}$, and α'' is the angle between mean normal $\overline{\mathbf{n}}_{ij}'^2$ and $\mathbf{l}$. Basically this operator gives the largest value to regions, where normals are oriented completely opposite to each other and the smallest value to regions, where normals have the same orientation (i.e. flat surfaces).

Ideally the factors $\triangle D_{ij}$, ω_2 and ω_2 should be calculated on each pair of opposite points and then summed up after multiplication. Due to small errors in the calculation of normals this approach is not very robust. Moreover, it is computationally expensive. Using only the mean points and normals to represent subpatches is a common approximation which proved to be accurate enough for our computations.



(a) (b) (c) (d)

Figure 2.4: Examples of 3D symmetry-based saliency maps calculated on artificial data. In the first row artificially created depth images of a cone, a rotated cube, a rotated cylinder and a sphere are shown in columns (a), (b), (c), (d) respectively. The second row shows the corresponding 3D symmetry-based maps calculated using kernel size $k = 30$.

The amount of 3D symmetry $S_{SYM3D}(r_i, c_i)$ at a given pixel $\mathbf{r}_i$ is equal to:

$$S_{SYM3D}(r_i, c_i) = \begin{cases} 0 & \text{if } D(r_i, c_i) = 0 \\ \max_{i=1,2,3}\{\Omega_i(\mathbf{r}_i)\} & \text{if } D(r_i, c_i) > 0 \end{cases} \tag{2.17}$$

where $D(r_i, c_i) = 0$ means that no depth information is available at this point.

Due to the nature of the 3D symmetry operator, convex and concave regions will obtain the same symmetry values. While in everyday scenarios the majority of objects that are claimed to be symmetric by humans, are rarely concave. To eliminate concave regions the following equation is applied:

$$S_{SYM3D}(r_i, c_i) = \begin{cases} 0 & \text{if } cos(\alpha')cos(\alpha'') < 0 \\ S_{SYM3D}(r_i, c_i) & \text{otherwise} \end{cases} \tag{2.18}$$

The 3D symmetry-based saliency map $SYM3D$ is linearly normalized to be in the range $[0, 1]$.

3D Symmetry-based Saliency Map on Artificial Data

To prove that our 3D symmetry operator is performing as expected, we have tested it on artificially created data. Artificially created data was produced from rendering mathematical models of different objects with known shape (i.e. cylinders, cubes, spheres, cones). Results of the symmetry operator are shown in Fig. 2.4.

From the presented results it is clearly visible that the proposed method works very accurately for synthetic examples. However, the result for the sphere (Fig. 2.4, (d)) visually does not look perfect, due to artifacts of the visualization process. Symmetry values for the

sphere are quite small (note that, as explained above, surface patches, that are rather flat locally, result in small values of ω_2). For visualization in Fig. 2.4, values were normalized to a visible range, which in this case led to an amplification of small errors from the normal calculation step.

2.4 Multi-Scale Saliency Maps

Operations involved in saliency calculation, such as normal estimation, depend on the size of the neighborhood or the kernel size and therefore are scale dependent. To avoid this scale dependency the saliency map is typically calculated on a multiple-scale Gaussian pyramid. The calculation on different scales allows to detect salient objects in different sizes in a computationally effective manner. In this Section, we describe how the Gaussian pyramid for the point cloud is created and then follow with a discussion of different approaches to calculate feature maps and across-scale combinations to create conspicuity maps, and conclude with a discussion of how master saliency maps can be obtained from those conspicuity maps.

2.4.1 Gaussian Pyramid for Point Clouds

Given a point cloud $\mathbf{P}$ we create X, Y, Z, N_x, N_y, N_z, so that $\forall \mathbf{r}_i \in \mathbf{P}$:

$$X\left(r_i, c_i\right) = x_i \tag{2.19}$$

$$Y\left(r_i, c_i\right) = y_i \tag{2.20}$$

$$Z\left(r_i, c_i\right) = z_i \tag{2.21}$$

$$N_x\left(r_i, c_i\right) = n_{ix} \tag{2.22}$$

$$N_y\left(r_i, c_i\right) = n_{iy} \tag{2.23}$$

$$N_z\left(r_i, c_i\right) = n_{iz} \tag{2.24}$$

If the point $\mathbf{r}_i$ does not have a valid depth value the corresponding value of images $I \in \{X, Y, Z, N_x, N_y, N_z\}$ equals to zero $I\left(r_i, c_i\right) = 0$.

To create a Gaussian Pyramid with L levels from the image I, the image is recursively convolved with a Gaussian kernel and downsampled by skipping even rows and columns. Only points that contain valid depth information are smoothed.

To obtain the image at the next level I_{j+1} we apply a Gaussian filter with a kernel $g = [1, 4, 6, 4, 1] \cdot [1, 4, 6, 4, 1]^T$ to the image at the current level I_j, where $I \in \{X, Y, Z, N_x, N_y, N_z\}$:

$$I'_{j+1}(r_i, c_i) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} g(m+2, n+2) I_j(r_i - m, c_i - n) * sign(Z_j(r_i - m, c_i - n) > 0) \tag{2.25}$$

where $Z(r_i, c_i) > 0$ means that the point has a valid depth value (the filter is implemented as a separable filter).

The filtered result $I'_{j+1}(r_i, c_i)$ is then normalized with the normalization factor:

$$N_{I'_j}(r_i, c_i) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} g(m+2, n+2) * sign(Z_j(r_i - m, c_i - n) > 0) \tag{2.26}$$

Image I'_{j+1} is then downsampled by skipping even rows and columns to obtain pyramid level I_{j+1}.

Gaussian pyramid from the point cloud P_l consists of L levels, and $\forall \mathbf{r}_i^l \in P_l$:

$$\mathbf{r}_i^l = \left(r_i^l, c_i^l, \mathbf{p}_i^l, \mathbf{n}_i^l\right) \tag{2.27}$$

where

$$\mathbf{p}_i^l = \left(X_l\left(r_i^l, c_i^l\right), Y_l\left(r_i^l, c_i^l\right), Z_l\left(r_i^l, c_i^l\right)\right) \tag{2.28}$$

$$\mathbf{n}_i^l = \left(N_{x_l}\left(r_i^l, c_i^l\right), N_{y_l}\left(r_i^l, c_i^l\right), N_{z_l}\left(r_i^l, c_i^l\right)\right) \tag{2.29}$$

and $\left(r_i^l, c_i^l\right)$ are row and column coordinates of point $\mathbf{r}_i^l$ in level l.

2.4.2 Feature Pyramid from Gaussian Pyramid

Several approaches to calculate feature pyramids from a Gaussian pyramid have been proposed in literature. One of the approaches proposed by Itti *et al.* [Itti 98] involved center-surround difference. Center-surround difference $\ominus$ between a "center" fine scale c and a "surround" coarser scale s reflects the fact, that salient regions are those that are different from its surrounding. Given a Gaussian pyramid with response maps $\{I_j\}$ (such as intensity or color opponency) feature maps $\{F_{c,s}\}$ at different levels are computed according to:

$$F_{c,s} = |I_c \ominus I_s| \tag{2.30}$$

where $c \in \{2, 3, 4\}$, $s = c + \delta$, and $\delta \in \{3, 4\}$ as was proposed by Itti *et al.* [Itti 98]. It is worth mentioning that values of c and s can be modified according to the task at hand and the original image size.

Frintrop *et al.* [Frintrop 05b] argued that approximation of on-center-off-surround, responding strongly to bright regions on a dark background and off-center-on-surround difference, responding strongly to dark regions on a bright background with simple center-surround difference as proposed by Itti *et al.* [Itti 98] yields problems when many peaks of one type and only few of the other exist. In that case attention should be guided towards those few peaks, while in the model of Itti *et al.* [Itti 98] those two types of peaks are equivalent. Therefore, Frintrop *et al.* [Frintrop 05b] proposed to create two types of feature pyramids $\{F_{c,s}^{on-off}\}$ and $\{F_{c,s}^{off-on}\}$ from the same Gaussian pyramid with response maps $\{I_j\}$:

$$F_{c,s}^{on-off} = I_c - mean\left(I, s\right) \tag{2.31}$$

$$F_{c,s}^{off-on} = mean\left(I, s\right) - I_c \tag{2.32}$$

where $c \in \{2, 3, 4\}$, $mean(I, s)$ represents the average of the surrounding pixels for two different sizes of the surround $s \in \{3, 7\}$ pixels.

Another approach to create a feature pyramid is to directly extract features from each level of the Gaussian pyramid:

$$F_l = f\left(I_l\right) \tag{2.33}$$

where $f\left(\cdot\right)$ is a specific function to create feature map from the response.

This approach was used by Kootstra *et al.* [Kootstra 08], where symmetry feature maps were calculated for each level of the Gaussian pyramid. Frintrop *et al.* [Frintrop 05b] also suggested that orientation maps obtained from intensity images after applying Gabor filters can serve themselves as feature maps.

On the contrary, Itti *et al.* [Itti 98] used those orientation maps to create a Gaussian pyramid and then applied center-surround operator to calculate feature maps:

$$F_{c,s} = \left|O\left(I_c\right) \ominus O\left(I_s\right)\right| \tag{2.34}$$

where $O\left(\cdot\right)$ is a Gabor filter at a specific orientation.

In this thesis we exploit two types of feature pyramids: feature pyramid $\{F_l\}$ built directly on the Gaussian pyramid of the point cloud and feature pyramid $\{F_{c,s}\}$ built using center-surround operator applied to the Gaussian pyramid from saliency maps. We build a Gaussian pyramid $\{P_l\}$ with $L = 8$ levels from point cloud P. Given a set of saliency operators $S_{TYPE} \in \{S_{PH}, S_{SH}, S_{RSO}, S_{SYM3D}\}$ the feature pyramid $\{F_l^{TYPE}\}$ is constructed as:

$$F_l^{TYPE} = S_{TYPE}(P_l) \tag{2.35}$$

To build the feature pyramid $\{F_{c,s}^{TYPE}\}$ we use the previously built pyramid $\{F_l^{TYPE}\}$:

$$F_{c,s}^{TYPE} = \left|F_c \ominus F_s\right| \tag{2.36}$$

where $c \in \{2, 3, 4\}$, $s = c + \delta$, and $\delta \in \{3, 4\}$ as was proposed by Itti *et al.* [Itti 98].

2.4.3 Normalization of Feature maps

Before feature maps from the pyramid can be combined together to form a conspicuity map, those feature maps should be normalized. In this Section we describe different normalization operators.

The simplest way to get normalized feature map F_l^N is to apply linear normalization to the range $[0, 1]$:

$$N_{LIN}\left(F_l\left(r_i, c_i\right)\right) = \frac{F_l\left(r_i, c_i\right) - F_l^{min}}{F_l^{max} - F_l^{min}} \tag{2.37}$$

where F_l^{min} and F_l^{max} are global minimum and global maximum values of the map F_l respectively.

Itti *et al.* [Itti 98] proposed to normalize feature maps in a way that a small number of strong peaks of activity is locally promoted, while numerous comparable peak responses are globally suppressed. This idea is similar to non-maxima suppression scheme and consists of several steps. At first, the linearly normalized saliency map F_l^N is produced. Then,

the global maximum F_l^{max} and the average of its other local maxima $\overline{F}_l^{max}$ of the feature map are found. The normalized feature map F_l^{NMS} is finally obtained according to:

$$N_{NMS}\left(F_l\left(r_i, c_i\right)\right) = F_l\left(r_i, c_i\right)\left(F_l^{max} - \overline{F}_l^{max}\right)^2 \tag{2.38}$$

Later, Itti *et al.* [Itti 99] argued that the above mentioned scheme is not biologically plausible and is not robust to noise. Instead an iterative scheme based on convolution with Gaussian filter was proposed. We leave out the details of this normalization operator, because iterative normalization is time consuming and is not suitable for robotics applications.

Frintrop *et al.* [Frintrop 05b] proposed a different normalization scheme, where the feature map is divided by the square root of the number of the local maxima N_{LM} in a pre-specified range from the global maximum:

$$N_{NLM}\left(F_l\left(r_i, c_i\right)\right) = \frac{F_l\left(r_i, c_i\right)}{\sqrt{N_{LM}}} \tag{2.39}$$

Please note, that there exist many other normalization operators, and in this Section we concentrated only on those that are used for further evaluation.

2.4.4 Conspicuity Map from Feature Pyramid

After feature pyramids F_l with L levels are computed and normalized they are combined together into a conspicuity map C. There exist different ways to create a conspicuity map such as across-scale addition:

$$C^{SUM} = N\left(\bigoplus_{l=2}^{l=C} N\left(F_l\right)\right) \tag{2.40}$$

and maximum value over all levels:

$$C^{MAX} = N\left(\max_{l=1..L} N\left(F_l\right)\right) \tag{2.41}$$

where $N(\cdot) \in \{N_{LIN}, N_{NMS}, N_{NLM}\}$ is a normalization operator.

2.4.5 Master Saliency Map from Conspicuity Maps

Given a set of conspicuity maps, in our case $C_{TYPE} \in \{C_{PH}, C_{SH}, C_{RSO}, C_{COL}\}$ a master saliency map is created. C_{COL} is a 2D color-based saliency map. In our experiments, we used the saliency map proposed by Harel *et al.* [Harel 06], because it showed the best performance among different evaluated color-based saliency maps. We investigate two different ways to combine those conspicuity maps: linear combination and maximization.

Master saliency map S_{SUM} obtained by linear combination:

$$S_{SUM} = N\left(\sum_{type \in \{PH,SH,RSO,COL\}} \omega_{type}\, C_{type}\right) \tag{2.42}$$

where ω_{type} is the weight of the conspicuity map C_{type}. The sum of all weight must be equal to 1:

$$\sum_{type \in \{PH,SH,RSO,COL\}} \omega_{type} = 1 \tag{2.43}$$

and in our case all weight are equal.

Master saliency map S_{MAX} obtained by maximization:

$$S_{MAX} = N\left(\max_{type \in \{PH,SH,RSO,COL\}} C_{type}\right) \tag{2.44}$$

2.5 Attention Points Extraction Strategies

As was mentioned in Sec. 2.1.3 there are several ways to evaluate saliency maps. In this thesis, we follow the idea of evaluating saliency maps based on the number of correctly detected objects of interest in the scene. To do this, we extract attention points, sometimes also called fixation points.

One of the main challenges of attention points extraction mechanisms is the inability to guarantee their uniqueness. What we essentially want is a system to detect objects, or if that is not possible due to clutter or occlusion, we want to at least detect good initial object candidates. Therefore, attention points can be used as seed points for attention-driven segmentation [Mishra 11] or as fixations for further investigation of the region like zooming in or foveation. This approach for scene investigation is highly useful in such applications as robot navigation, robot localization and object classification. It can significantly reduce the search space and automatically point to interesting objects or areas, without exhaustive and computationally expensive exploration of the whole scene. We concentrate here on the area of understanding how and what attention points should be extracted from the saliency map to obtain good object candidates.

In this Section, we will describe different strategies employed to extract attention points from saliency maps. We are going to discuss the following strategies: Winner-Take-All (WTA [Lee 99]), Maximum Salient Region (*MSR* [Frintrop 06a]), and a novel approach called T-Junctions. The saliency map S is used as input to calculate attention points.

2.5.1 Winner-Take-All

Lee *et al.* [Lee 99] showed that a Winner-Take-All (WTA) neural network can be used to simulate humans' behavior of scene components prioritization while observing a scene. Excitatory neurons in the network are independent and receive input from a saliency map and each neuron excites its corresponding WTA neuron. All WTA neurons evolve independently of each other. The "winner" is the one that fires first (*i. e.* reaches threshold). This triggers a complete reset of all other WTA neurons. Attention points can be defined as points of location of the "winner" neuron.

Firing of the "winner" neuron is followed by a shift of the Focus of Attention (FOA) and a local inhibition. FOA is shifted to be at the location of the "winner" neuron. Input neurons are inhibited at the new location of the FOA. Local inhibition prevents returning the FOA to

the just attended location and allows the next most salient location to become the winner. This process is called "Inhibition of Return" (IOR) and is described in [Posner 84]. All time constants, conductances, and firing thresholds used in the WTA model implemented are the same as in [Itti 98]. Attention points are extracted from down-sampled saliency maps.

2.5.2 Most Salient Region

Frintrop [Frintrop 06a] proposed the detection of the Most Salient Region (*MSR*) for object detection. The most salient point (attention point) determines the most salient region and is the point with the maximum saliency value. Starting from the attention point (seed), the surrounding salient region is extracted by means of seeded region growing. *MSR* consists of all neighbors of the seed with saliency values that differ by at most 25% from the saliency value of the attention point.

The focus of attention (FOA) is directed to the *MSR*. After that the *MSR* is inhibited and the next *MSR* is computed. Though this way of attention point detection is less biologically plausible than one proposed by Lee *et al.* [Lee 99], Frintrop [Frintrop 06a] argued that equivalent results are achieved with fewer computational resources.

2.5.3 T-Junction Attention Points

We propose a new T-Junction (TJ) attention points extraction strategy for 3D symmetry-based saliency maps. The intuition behind this strategy is to detect essential points on the object, *i. e.* those that are intersection points for visible surfaces composing the object. This strategy requires saliency maps in which each object is detected as an individual connected component (*i. e.* S_{SYM3D} Fig. 2.3.4, 2.29).

To extract an attention point we first extract connected components $\{C_k\}$ from saliency map S. Then we calculate skeletons $\{T_k\}$ of these connected components $\{C_k\}$. Attention points are defined as T-junction points of multi-segment skeletons, or as mid-points in the case of simple skeletal line segments.

2.6 Evaluation and Results

We have two primary goals of the evaluation described in this Section. The first goal is to compare the performance of the proposed saliency maps and state-of-the-art methods with respect to the task of object detection. The second goal is to understand how different multi-level strategies, across-scale combination methods, normalization algorithms, and different extraction strategies influence the performance of the detection. What we essentially want to know is if more complicated and therefore more computationally expensive approaches result in a better performance than simpler ones. It has a significant importance for robotic systems, where algorithms have to be not only reliable, but also fast.

Evaluation was done with respect to two metrics described in Section 2.6.1: *Hit Ratio* and *Distance to Center*.

We evaluated several start-of-the-art saliency maps to understand how different operators perform with respect to the task of object detection. The following state-of-the-art saliency operators were evaluated (Sec. 2.6.2):

- Itti *et al.* (ITTI, [Itti 98])
- Harel *et al.* (HAREL, [Harel 06])
- Hou *et al.* (HOU, [Hou 07])
- Bruce *et al.* (BRUCE, [Bruce 09])
- Kootstra *et al.* (KOOTSTRA, [Kootstra 10b])
- Wang *et al.* (WANG; WANG+ITTI; WANG+HOU; WANG+BRUCE, [Wang 13])

We evaluated the proposed saliency operators to understand which of them are the most suitable for our task of object detection. We also evaluated different multi-level, combination, normalization and extraction strategies to see which of them are best to be used for our task. The following proposed saliency operators were evaluated:

- *Point-based Height Saliency Map (PH)* (Sec. 2.6.3)
- *Surface-based Height Saliency Map (SH)* (Sec. 2.6.4)
- *Relative Surface Orientation Saliency Map (RSO)* (Sec. 2.6.5)
- *3D Symmetry-based Saliency Map (SYM3D)* (Sec. 2.6.6)

We also evaluated the combination (Sec. (2.4.5)) of the proposed saliency operators (Sec. 2.6.7) to understand if their combination leads to a better detection performance. Moreover, HAREL saliency map [Harel 06] was also used as an additional channel to create a master saliency map.

In total, 159 combinations between different saliency maps and extraction strategies were evaluated.

The experiments were performed on the *Object Segmentation Database (OSD)* [Richtsfeld 13]. The database consists of different table scenes including free-standing objects, multiple occluded objects and piles of objects. Objects in the database were hand-labeled with polygon masks. Examples of images from the database and different saliency maps are shown in Sec. 2.6.2 – Sec. 2.6.7.

Complete evaluation results are shown in Tables 2.4 and 2.5.

2.6.1 Evaluation Metrics

The comparison of different strategies to extract attention points applied to different types of saliency maps was done with respect to two metrics. The first metric is the *Hit Ratio* (*HR*), and the second metric is the *Distance to Center* (*DC*).

Hit Ratio

Hit Ratio (*HR*) shows how many different objects were covered by a given number of attention points, and is calculated as the percentage of unique attention points being situated inside different objects:

$$HR = \frac{n}{N} \tag{2.45}$$

where N is the number of attention points and n is the number of different attended objects. The higher (*HR*) the better is the attention mechanism.

Distance to Center

The distance between the extracted attention point (fixation point) f and the center of the respective object c is defined as:

$$DC(f, c) = \frac{\parallel f - c \parallel}{\underset{p \in object}{max} \parallel p - c \parallel} \tag{2.46}$$

The centers of the respective objects represent physical centers of the visible parts of the objects. The *DC* shows the accuracy of attention points. The smaller the distance, the better is the detection quality of attention points. A good attention point for an object is the one that is situated closer to the center of the object.

A perfect attention mechanism will hit every object exactly once, resulting in *HR* equaled to one and directly at the center, resulting in *DC* equaled to zero.

2.6.2 State-of-the-art Saliency Maps

In this Section we present evaluation results of the state-of-the-art saliency algorithms. For each type of the saliency map attention points for object detection were extracted using two different strategies: (1) Winner-Take-All (WTA, Sec. 2.5.1) and (2) Most Salient Region (MSR, Sec. 2.5.2).

The primary goal of the evaluation is to understand which type of saliency map and which extraction strategy is the most suitable for the task of object detection. This is also needed for the further evaluation of the proposed saliency maps, to understand if 3D information improves detection performance.

Hit Ratio

The evaluation (Fig. 2.5) shows that the best performance in terms of the Hit Ratio is achieved by the HAREL saliency map [Harel 06] both when MSR or WTA extraction strategies are applied. The worst performance in terms of the Hit Ratio is shown by HOU saliency map [Hou 07]. This means that saliency maps based on Feature Integration Theory (FIT) by Treisman *et al.* [Treisman 80] are more preferable for object detection, than information-based attentional models. This conclusion is also supported by low detection results for BRUCE saliency maps ([Bruce 09]).

ITTI saliency map [Itti 98] has worse performance than HAREL saliency map, despite the same underlaying principles. This finding can be explained by the fact that the two algorithms have different strategies for feature combination and normalization. It means that for color-based saliency maps the internal mechanism of saliency map computation is of high importance.

The performance of the KOOTSTRA saliency map [Kootstra 10b] is comparable with HAREL saliency map, however, is still lower. This shows that only 2D symmetry is not sufficient enough to detect objects in cluttered environments.

The saliency maps proposed by Wang *et al.* (WANG, WANG+HOU, WANG+ITTI, WANG+BRUCE) [Wang 13] and created using findings from human eye-tracking experiments in 3D, did not result in a boost of a performance in terms of object detection. At best, its performance (WANG+BRUCE) is similar to the HAREL saliency map, when WTA extraction strategy is used. This clearly shows, that there is a long way to go before research on visual attention can completely explain how humans search for objects in 3D space and obtained eye-tracking results can be successfully used in robotics applications.

The fact that the performance of all saliency maps drops significantly with the number of extracted attention points, means that multiple attention points discover the same object. However, our ultimate goal is to discover each object precisely once. Therefore, a better attentional operator and extraction mechanism must be employed. Evaluation results of the proposed saliency maps, that are described below, show the benefits of the proposed saliency maps with respect to the detection task.

Distance to Center

In terms of the Distance to Center, symmetry-based saliency map KOOTSTRA shows the best performance when WTA extraction strategy is used. HAREL saliency map shows the worst performance with MSR extraction strategy. Distance to the Center does not change significantly neither with the increase of the number of attention points, nor when different saliency maps or extraction mechanisms are used.

Results

In general the performance of all state-of-the-art saliency operators are comparable, within standard deviations and there is no clear benefit of using one preferred saliency map or extraction strategy. Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.6, Fig. 2.7, Fig. 2.8.

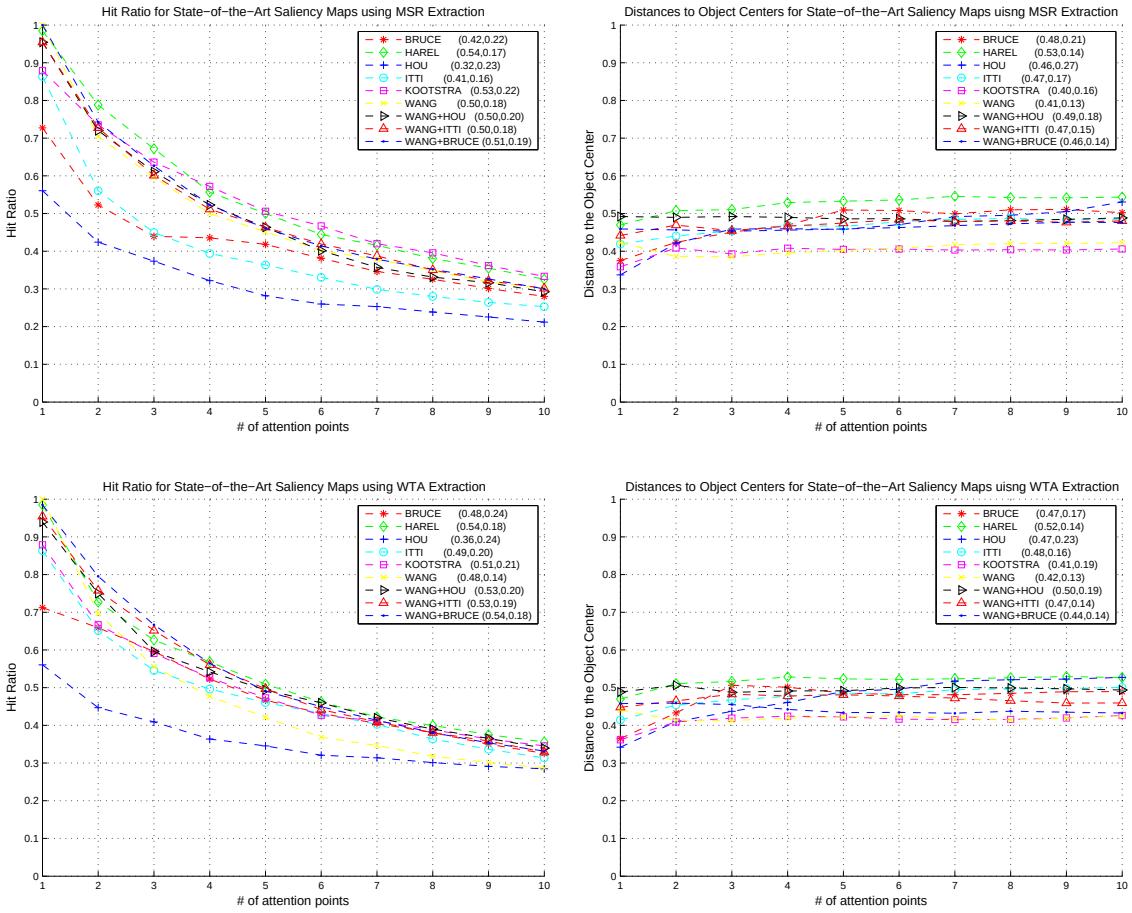
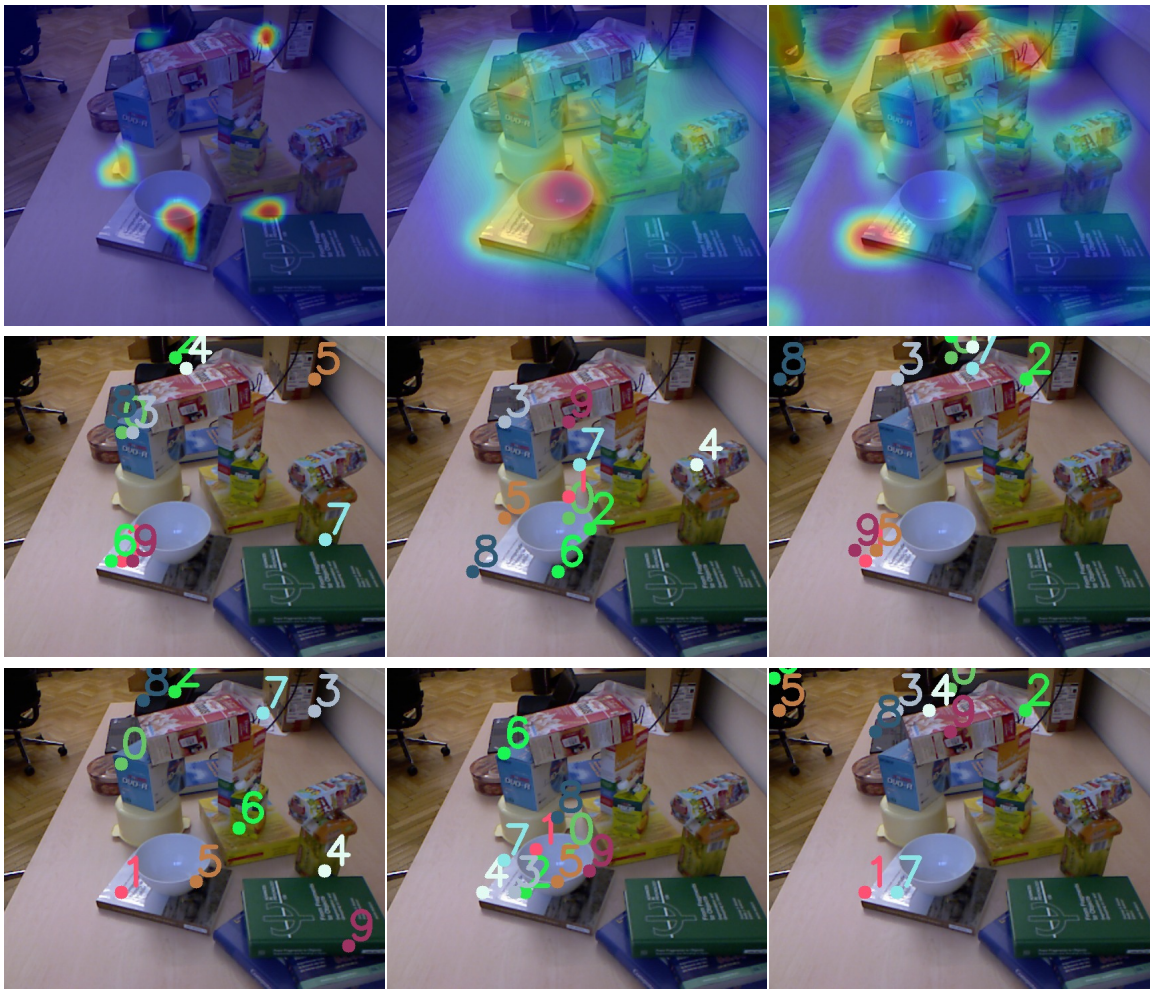
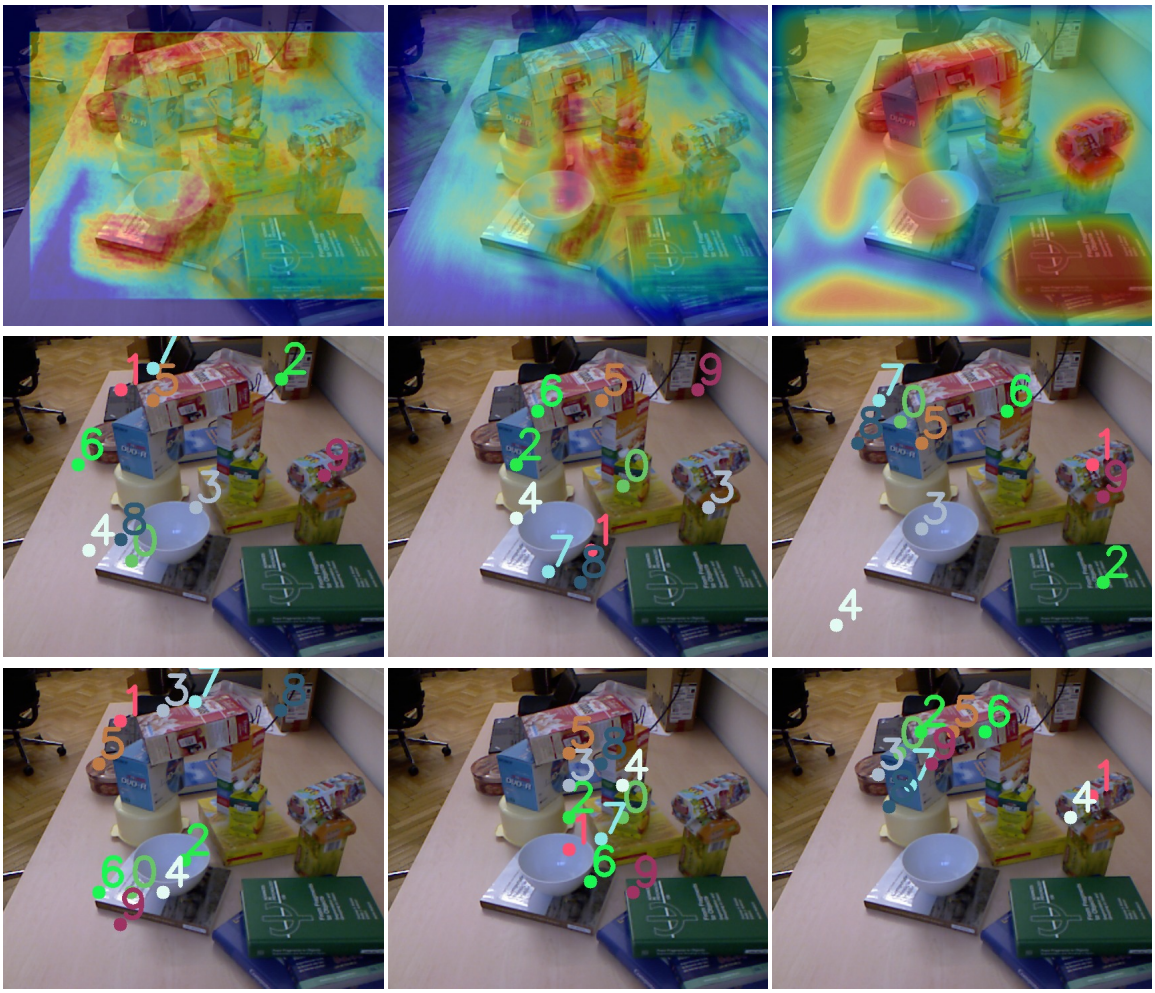


Figure 2.5: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for state-of-the-art saliency maps. Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color. (Numbers in brackets represent average values and standard deviation.)



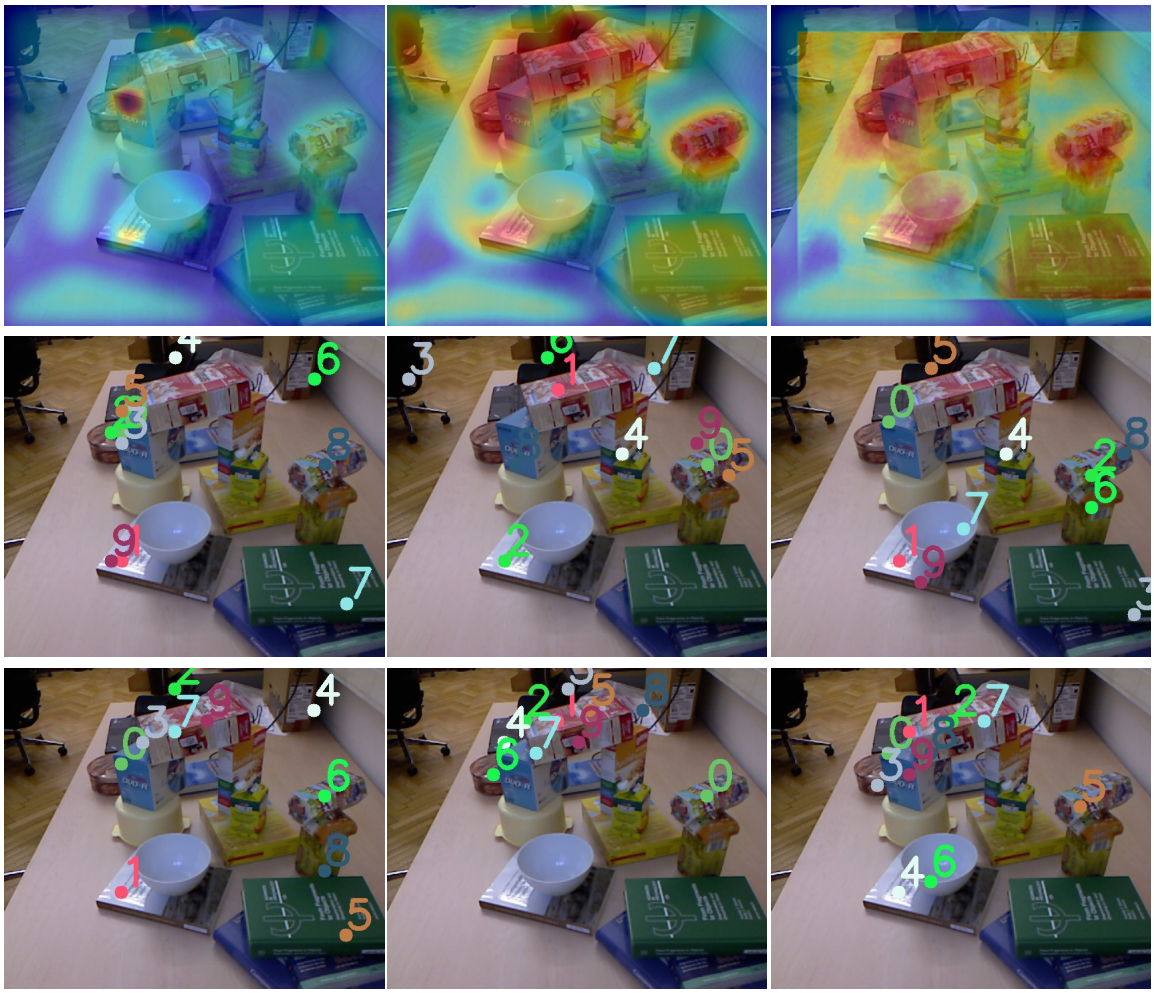
ITTI ([Itti 98]) HAREL ([Harel 06]) HOU ([Hou 07])

Figure 2.6: Row 1 shows examples of ITTI ([Itti 98]), HAREL ([Harel 06]) and HOU ([Hou 07]) saliency maps respectively overlaid with images from the Object Segmentation Database (OSD). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



BRUCE ([Bruce 09]) KOOTSTRA ([Kootstra 10b]) WANG ([Wang 13])

Figure 2.7: Row 1 shows examples of BRUCE ([Bruce 09]), KOOTSTRA ([Kootstra 10b]) and WANG ([Wang 13]) saliency maps respectively overlaid with images from the Object Segmentation Database (OSD). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



WANG+ITTI ([Wang 13]) WANG+HOU ([Wang 13]) WANG+BRUCE ([Wang 13])

Figure 2.8: Row 1 shows examples of WANG+ITTI, WANG+HOU and WANG+BRUCE ([Wang 13]) saliency maps respectively overlaid with images from the Object Segmentation Database (OSD). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note, that pictures are best seen in color.

2.6.3 Point-based Height Saliency Maps

In this Section we present evaluation results for Point-based Height saliency maps (PH, Sec. 2.3.1). We used the following multi-level approaches to create feature maps (Sec. 2.4.2): (1) Simple Feature Pyramid $\{F_l\}$ (SIMPLE) and (2) Itti-based Feature Pyramid $\{F_{c,s}\}$ (ITTI).

Combination of levels in the feature pyramid was done using two types of techniques (Sec. 2.4.4): (1) across-scale addition (SUM) and (2) maximization (MAX).

Normalization of saliency maps was done using the following normalization operators (Sec. 2.4.3): (1) linear normalization (LIN), (2) non-maxima suppression (NMS) and (3) non-linear maximization (NLM).

Attention points for object detection were extracted using two different methods: (1) Winner-Take-All (WTA, Sec. 2.5.1), (2) Most Salient Region (MSR, Sec. 2.5.2).

Our goal in this Section is to understand which multi-level approach, combination and normalization methods along with extraction strategy are the most suitable for Point-based Height saliency maps to be used in object detection. We also want to understand if more complicated strategies result in increased performance.

Multi-level Point-based Height Saliency Maps using Simple Feature Pyramid

Hit Ratio

The evaluation of multi-level Point-based Height saliency maps (Fig. 2.9) using Simple Feature Pyramid (SIMPLE) shows that the best performance in terms of the Hit Ratio is achieved when conspicuity maps on different levels are combined using across-scale addition (SUM) with linear normalization (LIN) and MSR extraction strategy are applied. The worst performance in terms of the Hit Ratio was shown when MSR extraction strategy is applied to saliency map based on maximization (MAX) with linear normalization (LIN).

In general, when MSR extraction strategy is used, across-scale combination (SUM) gives better results than combination by maximization (MAX) with no respect to a normalization method. WTA extraction strategy resulted in lower performance compared to MSR, regardless of which combination and normalization types were used.

Distance to Center

The Distance to Center metric gives the best result and the worst results when MSR and WTA extraction strategies are applied respectively to the saliency map obtained using combination by maximization (MAX) and linear normalization (LIN). Moreover, the overall performance of different variants of Point Height saliency maps is better when MSR extraction strategy is used. This potentially means that for robotics applications the simplest, and therefore the fastest, strategy must be selected, since there is no apparent improvements when more sophisticated mechanisms are used.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.10, Fig. 2.11, Fig. 2.12.

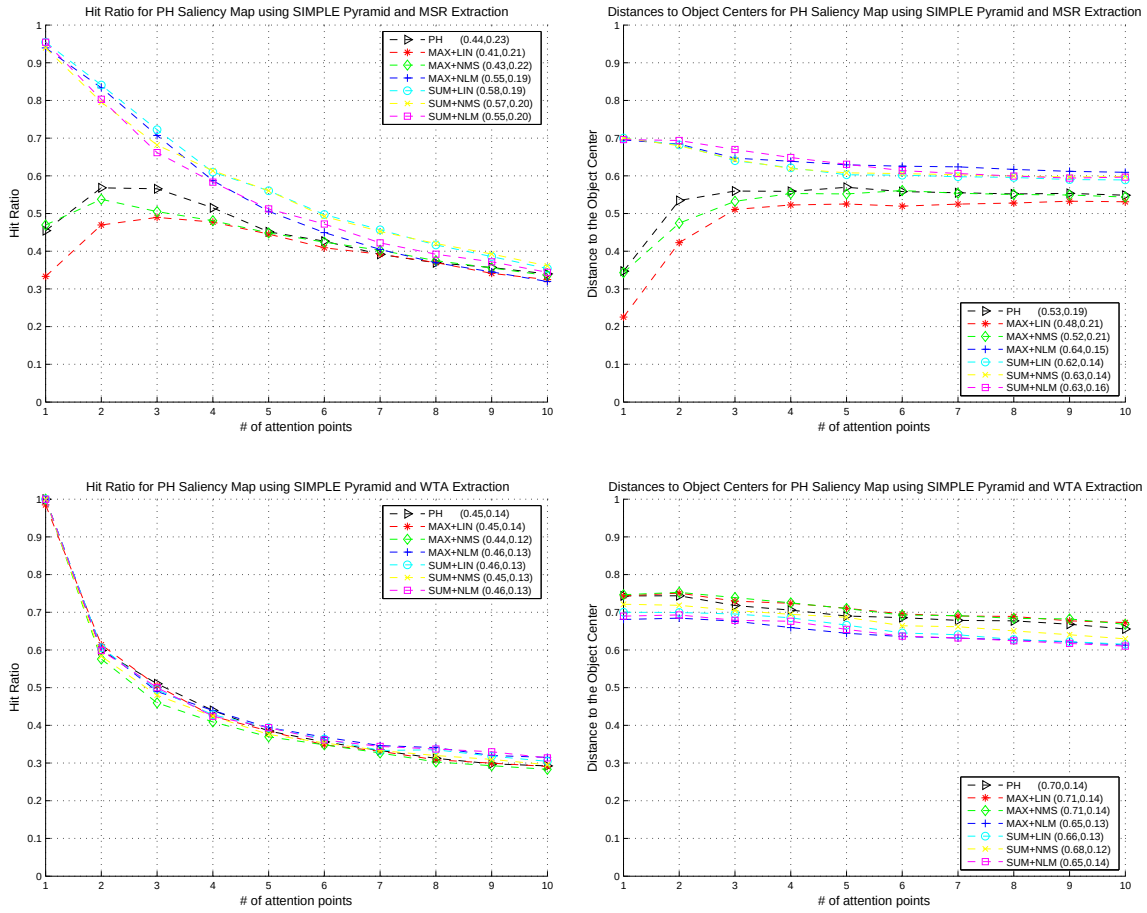
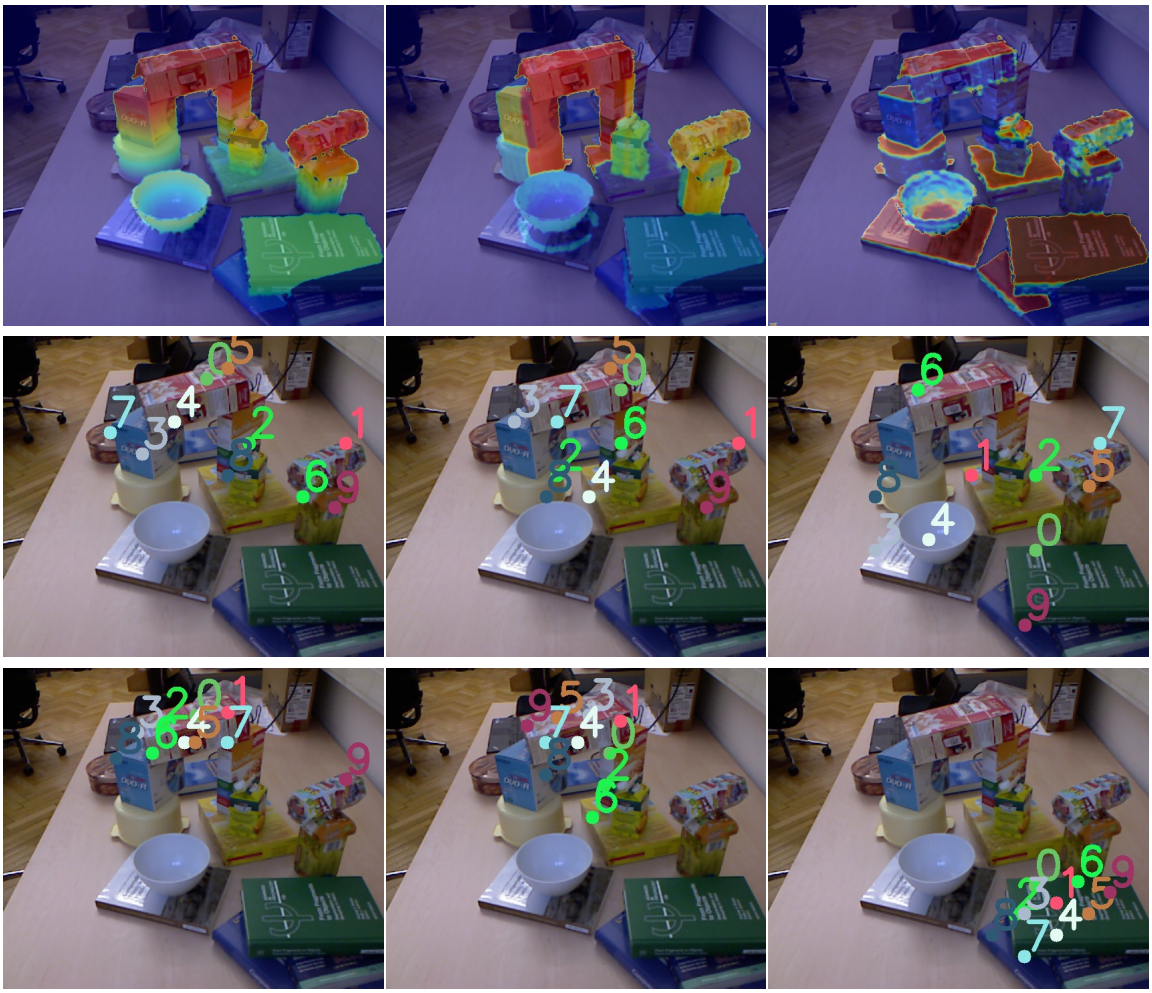
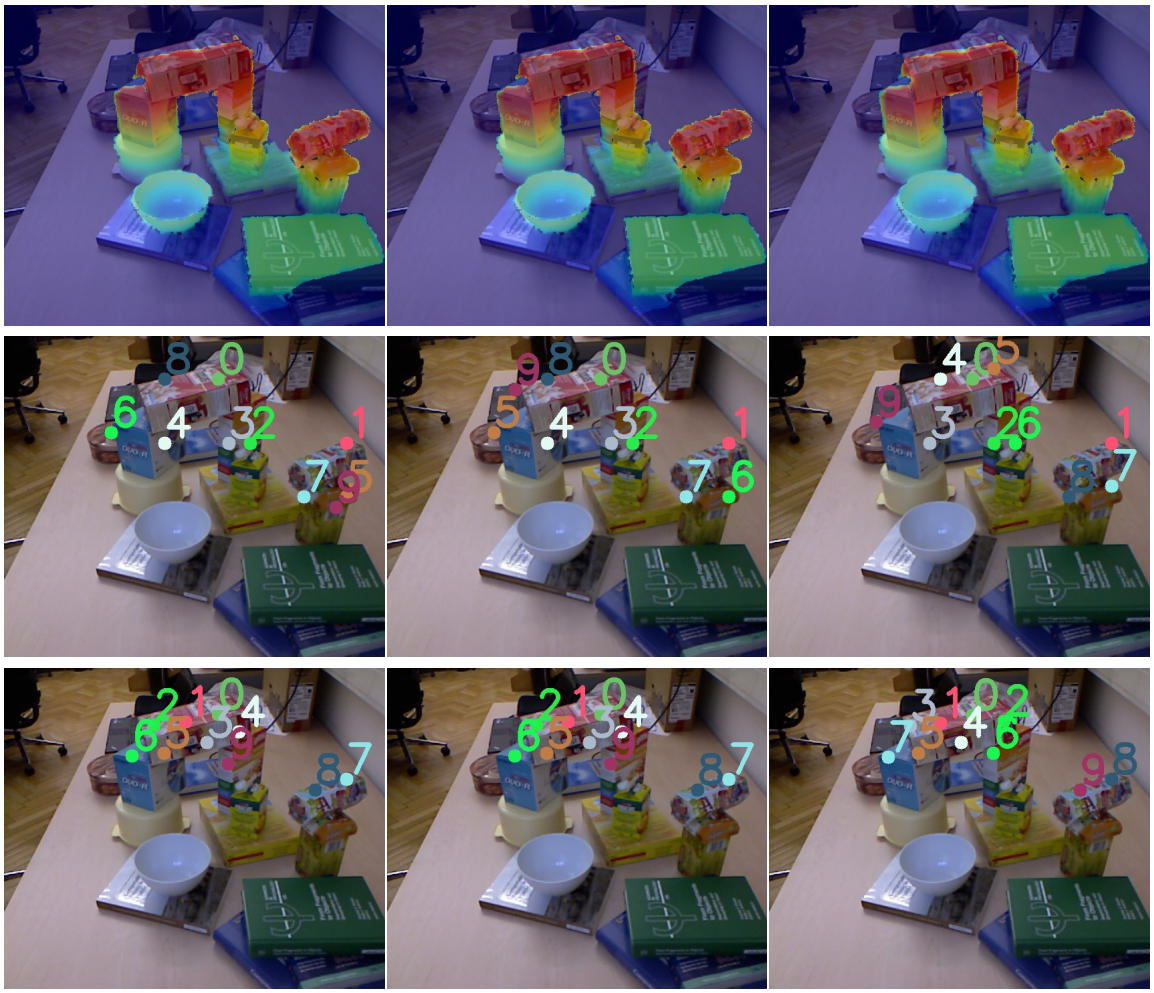


Figure 2.9: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Point Height saliency map (PH, Sec. 2.3.1) calculated using $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color. (Numbers in brackets represent average values and standard deviation.)



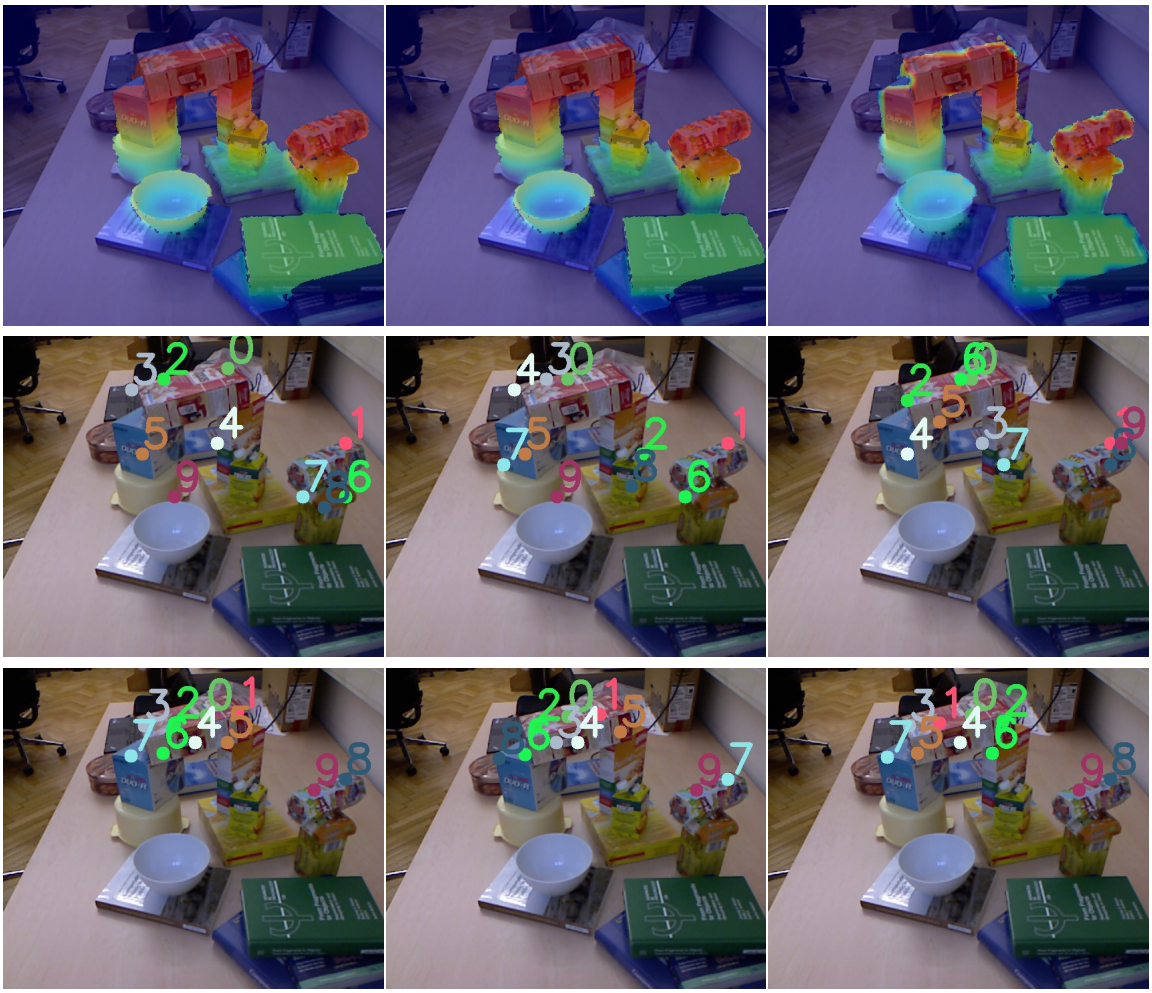
PH SH RSO

Figure 2.10: Row 1 shows examples of one level (SINGLE) Point Height saliency map (PH, Sec. 2.3.1), one level (SINGLE) Surface Height saliency map (SH, Sec. 2.3.2), and one level (SINGLE) Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) overlaid with images from the Object Segmentation Database (OSD). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



SIMPLE + SUM + LIN SIMPLE + SUM + NMS SIMPLE + SUM + NLM

Figure 2.11: Row 1 shows examples of Point Height saliency map (PH, Sec. 2.3.1) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



SIMPLE + MAX + LIN SIMPLE + MAX + NMS SIMPLE + MAX + NLM

Figure 2.12: Row 1 shows examples of Point Height saliency map (PH, Sec. 2.3.1) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

Multi-level Point-based Height Saliency Maps using Itti-based Feature Pyramid

Hit Ratio

The evaluation of multi-level Point-based Height saliency maps (Fig. 2.13) using Itti-based Feature Pyramid (ITTI) shows that the best performance in terms of the Hit Ratio is achieved when conspicuity maps on different levels are combined using maximization (MAX) with linear normalization (LIN), and MSR extraction strategy is applied. However, if normalization is changed to non-maxima suppression (NMS), the performance drops to the worst. This shows that for Itti-based feature pyramid, normalization is an essential part and must be carefully considered. In general there are no clear benefits of using one or the other combination or normalization type, except that when WTA strategy is applied, combination by across-scale addition gives better results on an average.

Distance to Center

The Distance to Center metric shows better performance, when MSR extraction strategy is applied. Moreover, a one level Point Height saliency map gives better detection accuracy, than any other variation. The worst results are shown when across-scale summation (SUM) with non-linear maximization normalization (NLM) and WTA extraction strategy are used. In general, all different combination types give similar results within the standard deviation.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.14, Fig. 2.15.

Results

Comparing evaluation results obtained using SIMPLE and ITTI pyramids, we can conclude that no clear improvements in detection quality were found when a more complicated scheme is involved. This means that for Point-based Height saliency maps a simpler approach can be used without any loss in detection results.

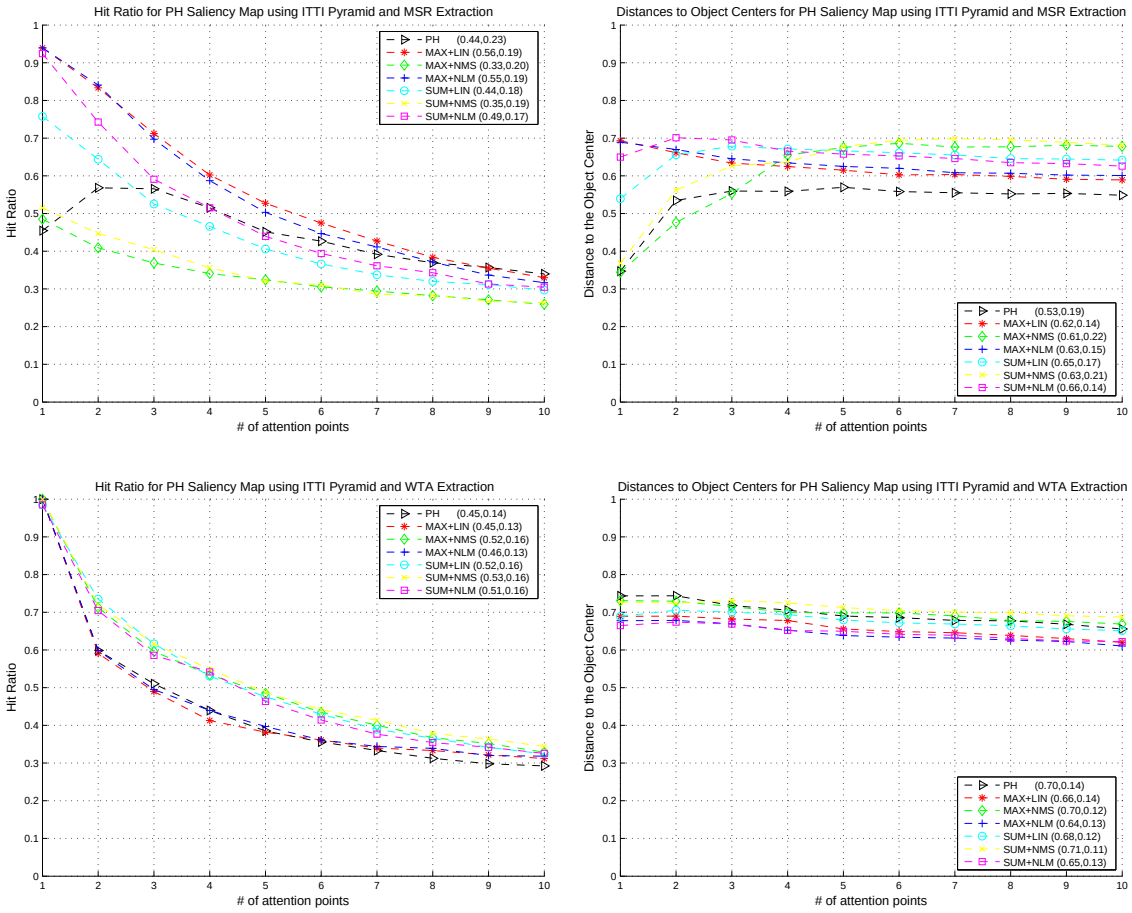
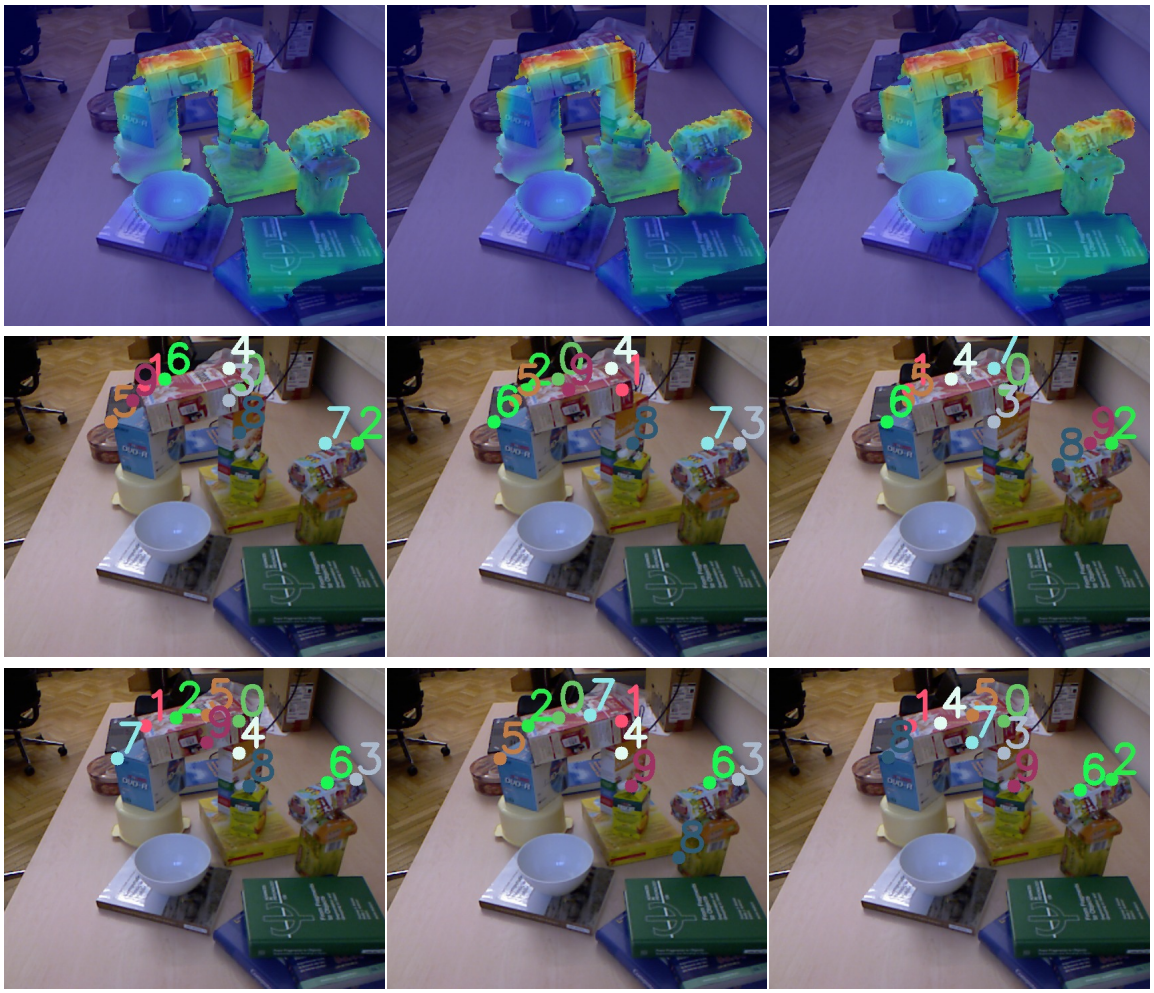
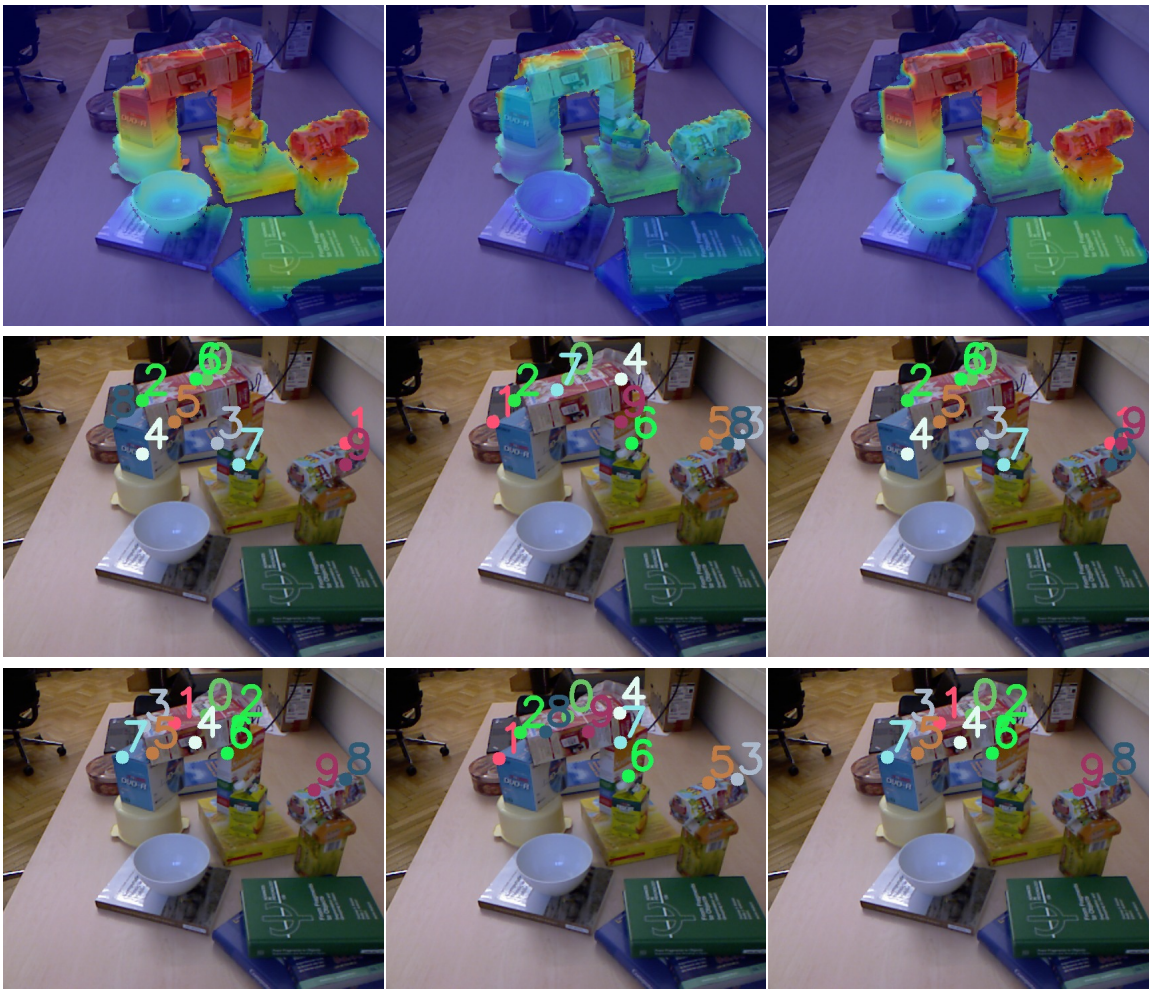


Figure 2.13: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Point Height saliency map (PH, Sec. 2.3.1) calculated using $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color.



ITTI + SUM + LIN ITTI + SUM + NMS ITTI + SUM + NLM

Figure 2.14: Row 1 shows examples of Point Height saliency map (PH, Sec. 2.3.1) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



ITTI + MAX + LIN ITTI + MAX + NMS ITTI + MAX + NLM

Figure 2.15: Row 1 shows examples of Point Height saliency map (PH, Sec. 2.3.1) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

2.6.4 Surface-based Height Saliency Maps

In this Section we present evaluation results for Surface-based Height saliency maps (SH, Sec. 2.3.2). We used the following multi-level approaches to create feature maps (Sec. 2.4.2): (1) Simple Feature Pyramid $\{F_l\}$ (SIMPLE) and (2) Itti-based Feature Pyramid $\{F_{c,s}\}$ (ITTI).

Combination of levels in the feature pyramid was done using two types of techniques (Sec. 2.4.4): (1) across-scale addition (SUM) and (2) maximization (MAX).

Normalization of saliency maps was done using the following normalization operators (Sec. 2.4.3): (1) linear normalization (LIN), (2) non-maxima suppression (NMS) and (3) non-linear maximization (NLM).

Attention points for object detection were extracted using two different methods: (1) Winner-Take-All (WTA, Sec. 2.5.1), (2) Most Salient Region (MSR, Sec. 2.5.2).

Our goal in this Section is to understand what multi-level approach, combination and normalization methods along with extraction strategy are the most suitable for Surface-based Height saliency map when used for the object detection. We also want to understand if the usage of more sophisticated strategies results in increased performance.

Multi-level Surface-based Height Saliency using Simple Feature Pyramid

Hit Ratio

The evaluation of multi-level Surface-based Height saliency maps (Fig. 2.16) using Simple Feature Pyramid (SIMPLE) shows that the best performance in terms of Hit Ratio is achieved when conspicuity maps on different levels are combined using across-scale addition (SUM) with linear normalization (LIN), and MSR extraction strategy is applied. The worst performance in terms of the Hit Ratio is shown when WTA extraction strategy is applied to saliency maps created using combination by maximization (MAX) and either linear normalization (LIN) or non-maxima suppression normalization (NMS). Overall, the performance of MSR extraction strategy is better on an average by 30%–40% than WTA extraction strategy.

Distance to Center

In terms of the Distance to Center metric, maximization (MAX) with non-maxima suppression normalization (NMS) and MSR strategy gives the best result. If maximization (MAX) with linear normalization (LIN) and WTA extraction are used, then the Distance to Center metric shows the worst performance. In general, MSR shows better performance than WTA with respect to Distance to Center metric.

Overall, the performance of different modifications of Surface Height saliency map calculated using $\{F_l\}$ pyramid (SIMPLE) shows better detection performance when MSR extraction strategy is applied. Moreover, using across-scale summation of different pyramid levels with MSR extraction strategy gives better results than any other type of combination.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.10, Fig. 2.17, Fig. 2.18.

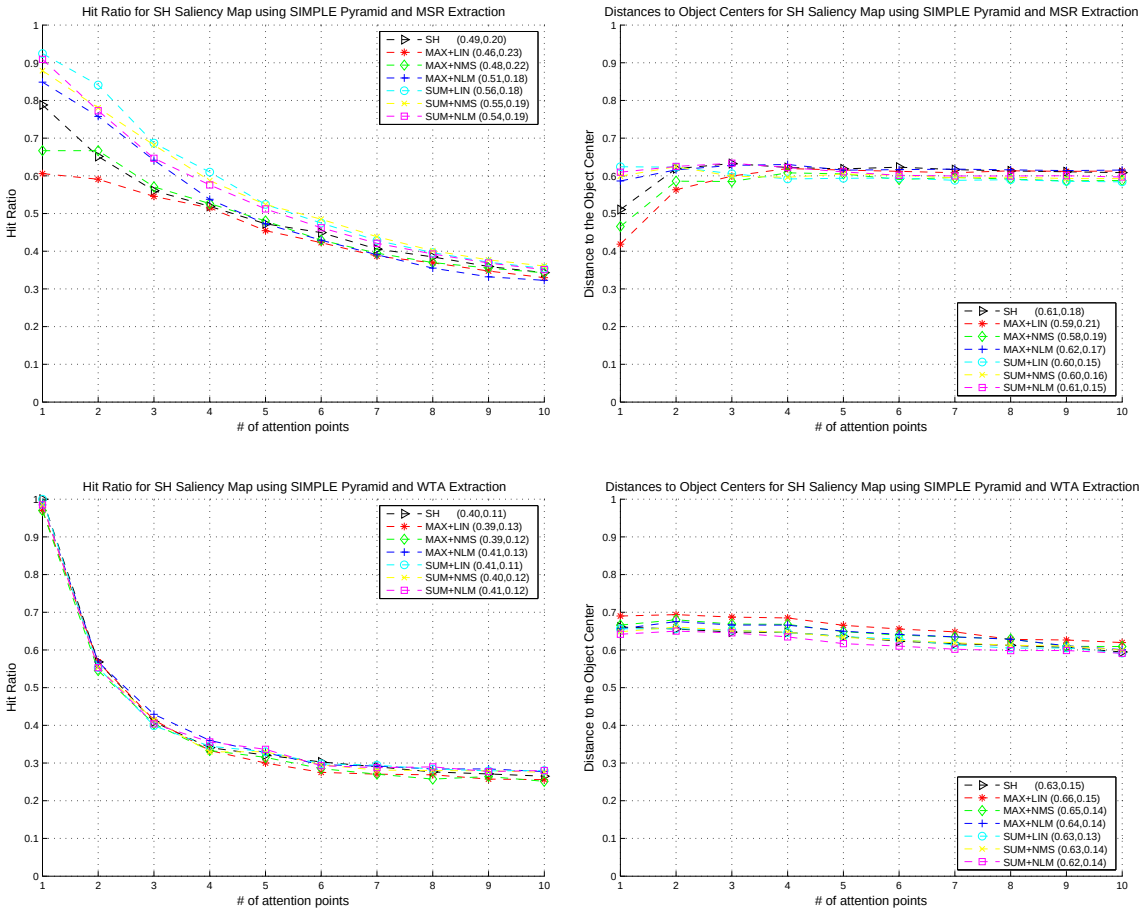
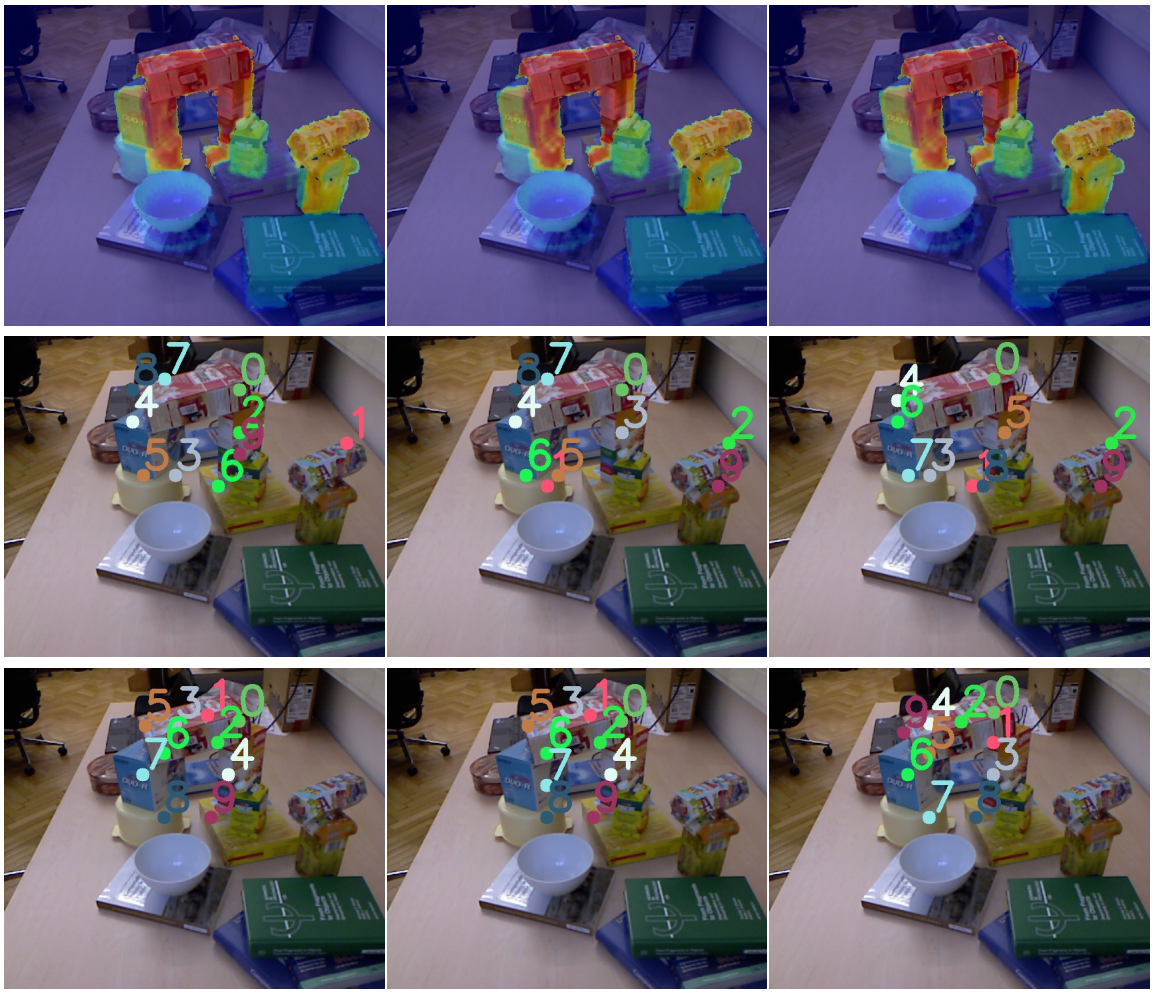
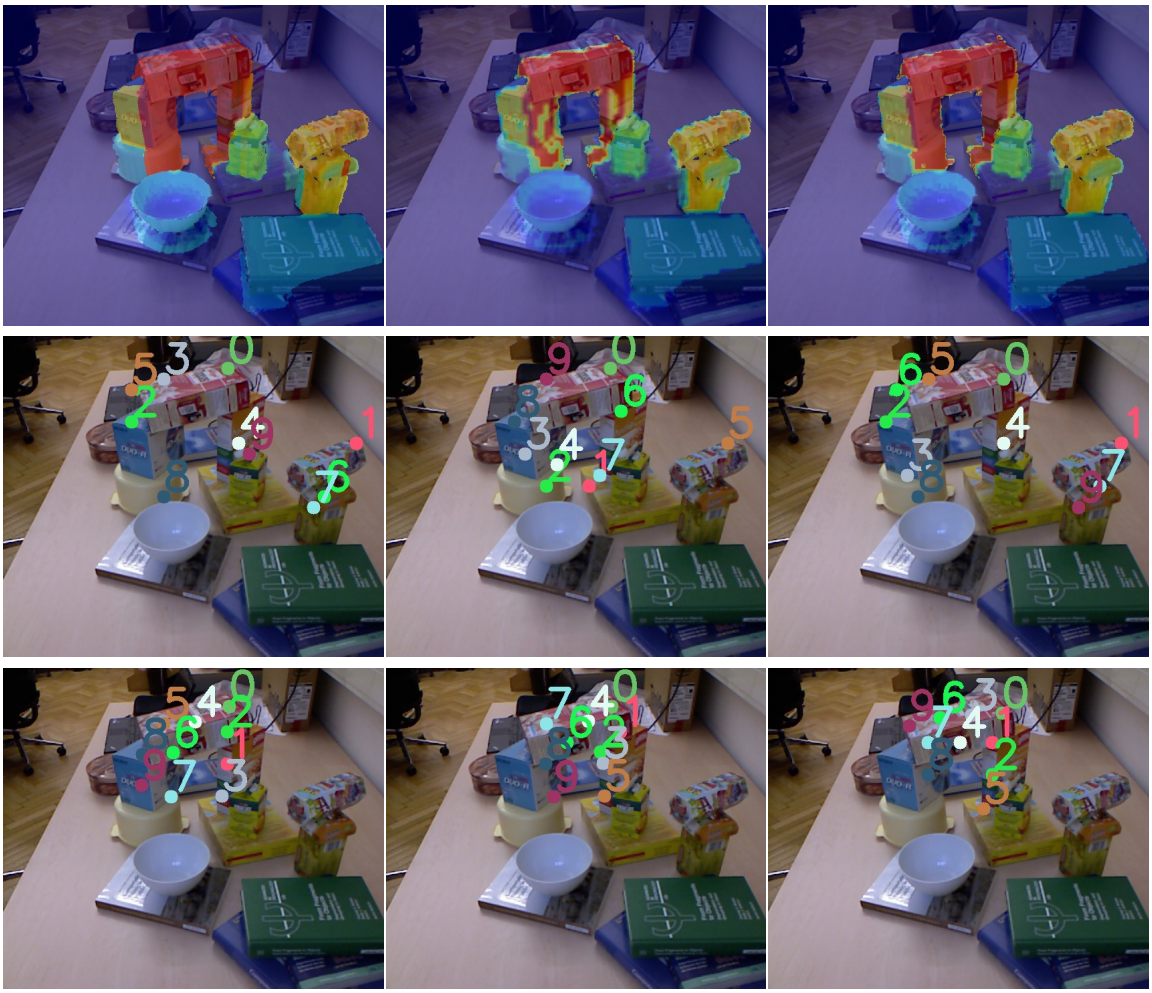


Figure 2.16: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Surface Height saliency map (SH, Sec. 2.3.2) calculated using $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color. (Numbers in brackets represent average values and standard deviation.)



SIMPLE + SUM + LIN SIMPLE + SUM + NMS SIMPLE + SUM + NLM

Figure 2.17: Row 1 shows examples of Surface Height saliency map (SH, Sec. 2.3.2) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



SIMPLE + MAX + LIN SIMPLE + MAX + NMS SIMPLE + MAX + NLM

Figure 2.18: Row 1 shows examples of Surface Height saliency map (SH, Sec. 2.3.2) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

Multi-level Surface-based Height Saliency Maps using Itti-based Feature Pyramid

Hit Ratio

The evaluation of multi-level Surface-based Height saliency maps (Fig. 2.19) using Itti-based Feature Pyramid (ITTI) shows that the best performance in terms of the Hit Ratio is achieved when conspicuity maps on different levels are combined using maximization (MAX) with linear normalization (LIN), and MSR extraction strategy is used. However, if normalization is changed to non-maxima suppression (NMS), then the performance drops to the worst. It worth mentioning that the performance of one-level saliency maps was almost identical to the best possible. This means that for this type of saliency maps using multi-level approach does not bring any improvements.

Distance to Center

For Surface-based Height saliency maps when combination using maximization (MAX) with non-linear maximization normalization (NLM), and MSR extraction strategy is used the Distance to the Center metric gives the best result. The worst result is achieved when across-scale summation (SUM) with non-maxima suppression (NMS) and WTA extraction strategy is used. Overall, evaluation results for Distance to Center metric are similar between different combination, normalization and extraction types.

Results

In general, results showed that different combination types showed similar results within standard deviations. No obvious benefits of using one specific type of combination, normalization or extraction strategy was found. Moreover, using ITTI feature pyramid type did not bring any improvements in detection performance. Therefore, the strategy of using the simplest possible combination, normalization and extraction are preferable.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.20, Fig. 2.21.

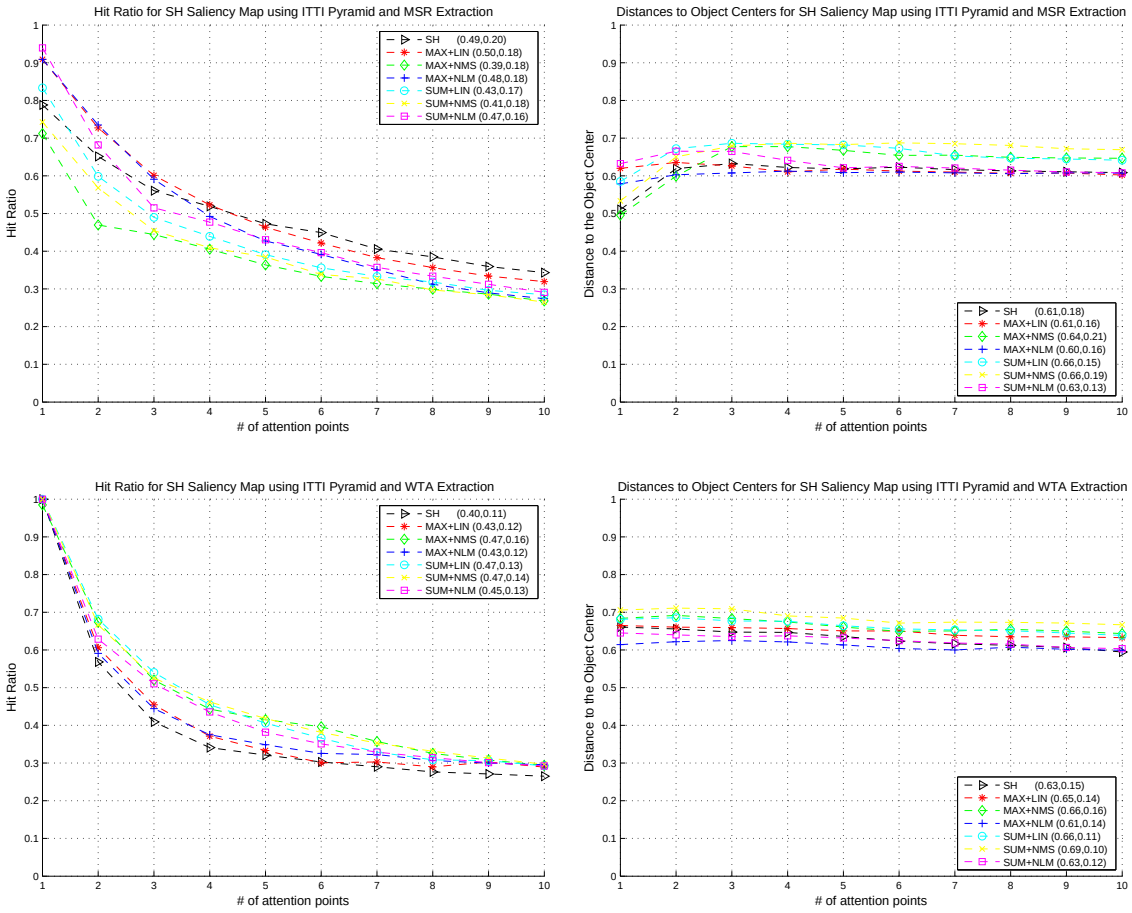
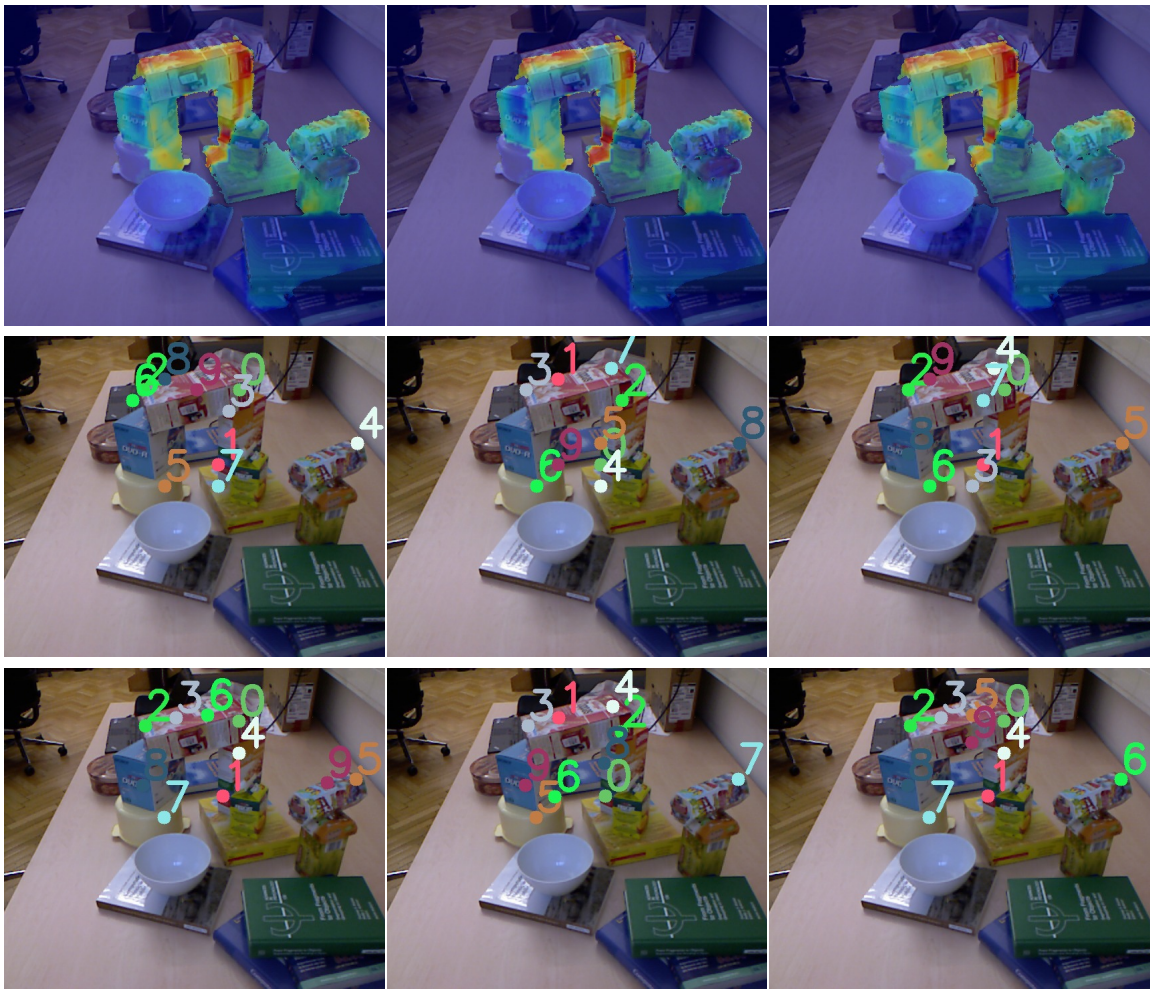
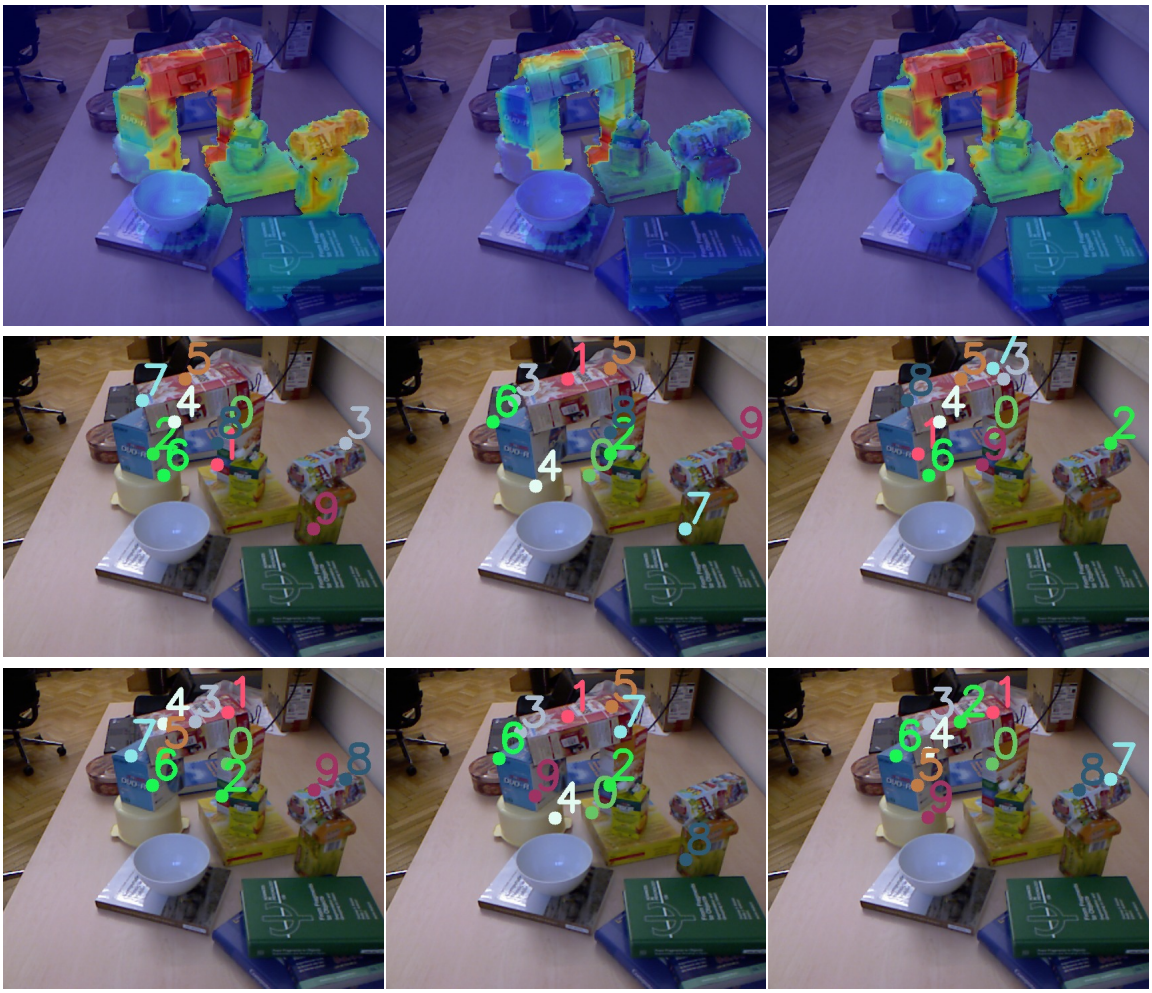


Figure 2.19: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Surface Height saliency map (SH, Sec. 2.3.2) calculated using $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color.



ITTI + SUM + LIN ITTI + SUM + NMS ITTI + SUM + NLM

Figure 2.20: Row 1 shows examples of Surface Height saliency map (SH, Sec. 2.3.2) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



ITTI + MAX + LIN ITTI + MAX + NMS ITTI + MAX + NLM

Figure 2.21: Row 1 shows examples of Surface Height saliency map (SH, Sec. 2.3.2) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

2.6.5 Relative Surface Orientation Saliency Maps

In this section we present evaluation results for Relative Surface Orientation saliency maps (RSO, Sec. 2.3.3). We used the following multi-level approaches to create feature maps (Sec. 2.4.2): (1) Simple Feature Pyramid $\{F_l\}$ (SIMPLE) and (2) Itti-based Feature Pyramid $\{F_{c,s}\}$ (ITTI).

Combination of levels in the feature pyramid was done using two types of techniques (Sec. 2.4.4): (1) across-scale addition (SUM) and (2) maximization (MAX).

Normalization of saliency maps was done using the following normalization operators (Sec. 2.4.3): (1) linear normalization (LIN), (2) non-maxima suppression (NMS) and (3) non-linear maximization (NLM).

Attention points for object detection were extracted using two different methods: (1) Winner-Take-All (WTA, Sec. 2.5.1), (2) Most Salient Region (MSR, Sec. 2.5.2).

Our goal in this Section is to understand what multi-level approach, combination and normalization methods along with extraction strategy are the most suitable for Relative Surface Orientation saliency maps to be used in object detection. We also want to understand if the usage of more complicated strategies results in increased performance.

Multi-level Relative Surface Orientation Saliency Maps using Simple Feature Pyramid

Hit Ratio

The evaluation of multi-level Relative Surface Orientation saliency maps (Fig. 2.22) using Simple Feature Pyramid shows that the best performance in terms of Hit Ratio is achieved when conspicuity maps on different levels are combined using across-scale addition (SUM) with linear normalization (LIN), and MSR extraction strategy is applied. The worst performance in terms of the Hit Ratio is shown when WTA extraction strategy is applied to saliency map based on maximization (MAX) with linear normalization (LIN). In general, MSR extraction strategy showed better performance than WTA extraction strategy regardless of the type of combination and normalization.

Distance to Center

In terms of the Distance to Center metric, across-scale addition (SUM) both with linear (LIN) or non-maxima suppression (NMS) normalizations and WTA strategy gives the best result. If maximization (MAX) with linear normalization (LIN) and WTA extraction is used, the Distance to Center metric shows the worst performance.

In general no specific combination or normalization type was superior to all others.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.10, Fig. 2.23, Fig. 2.24.

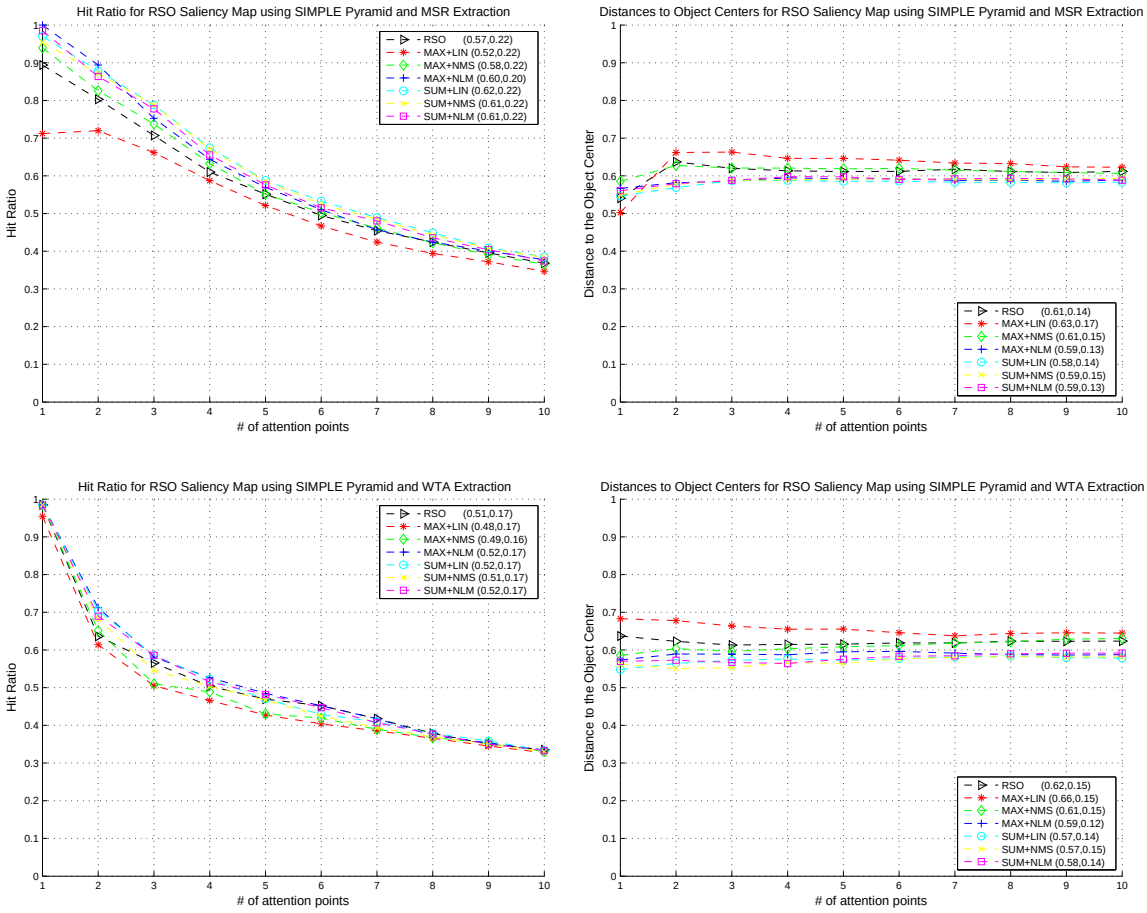
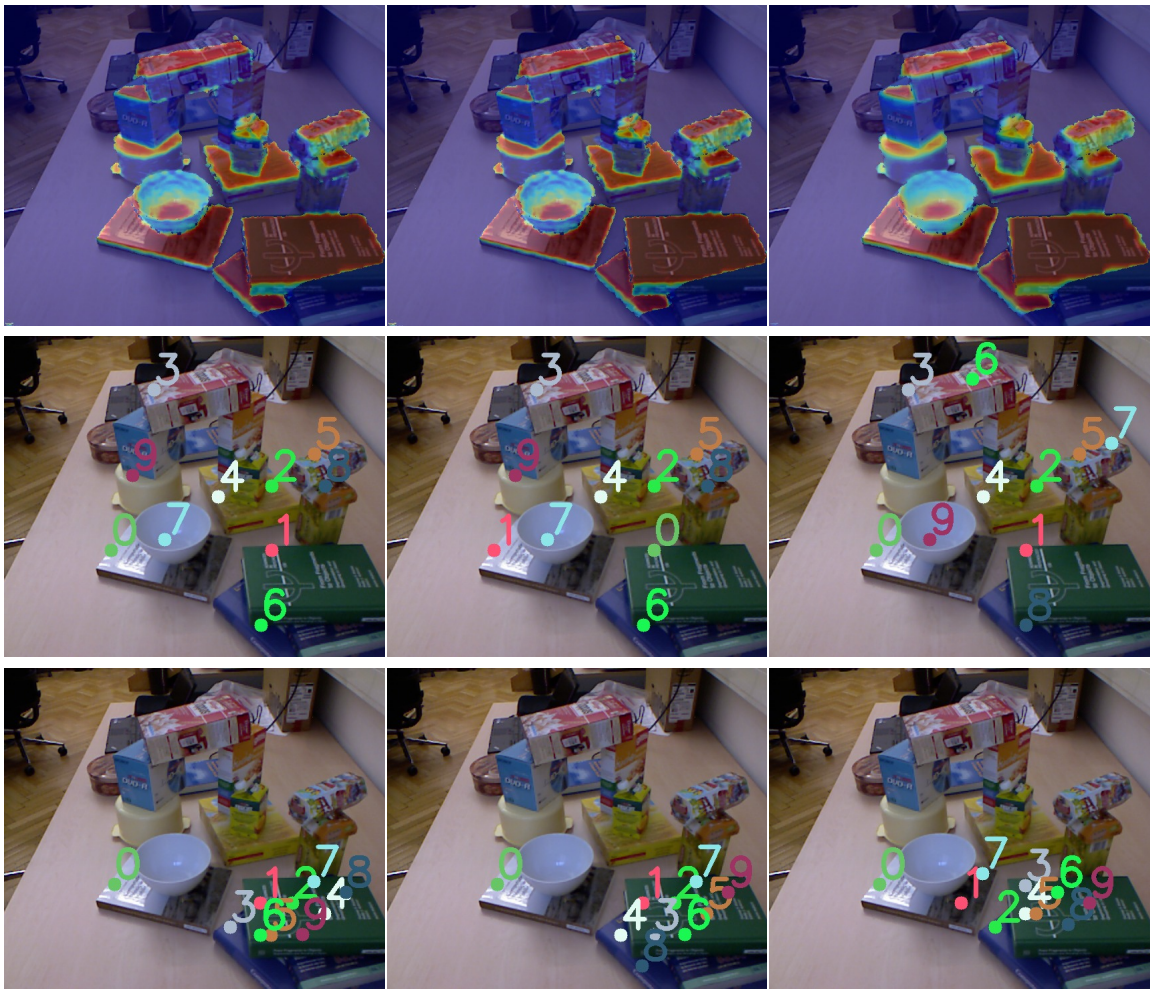
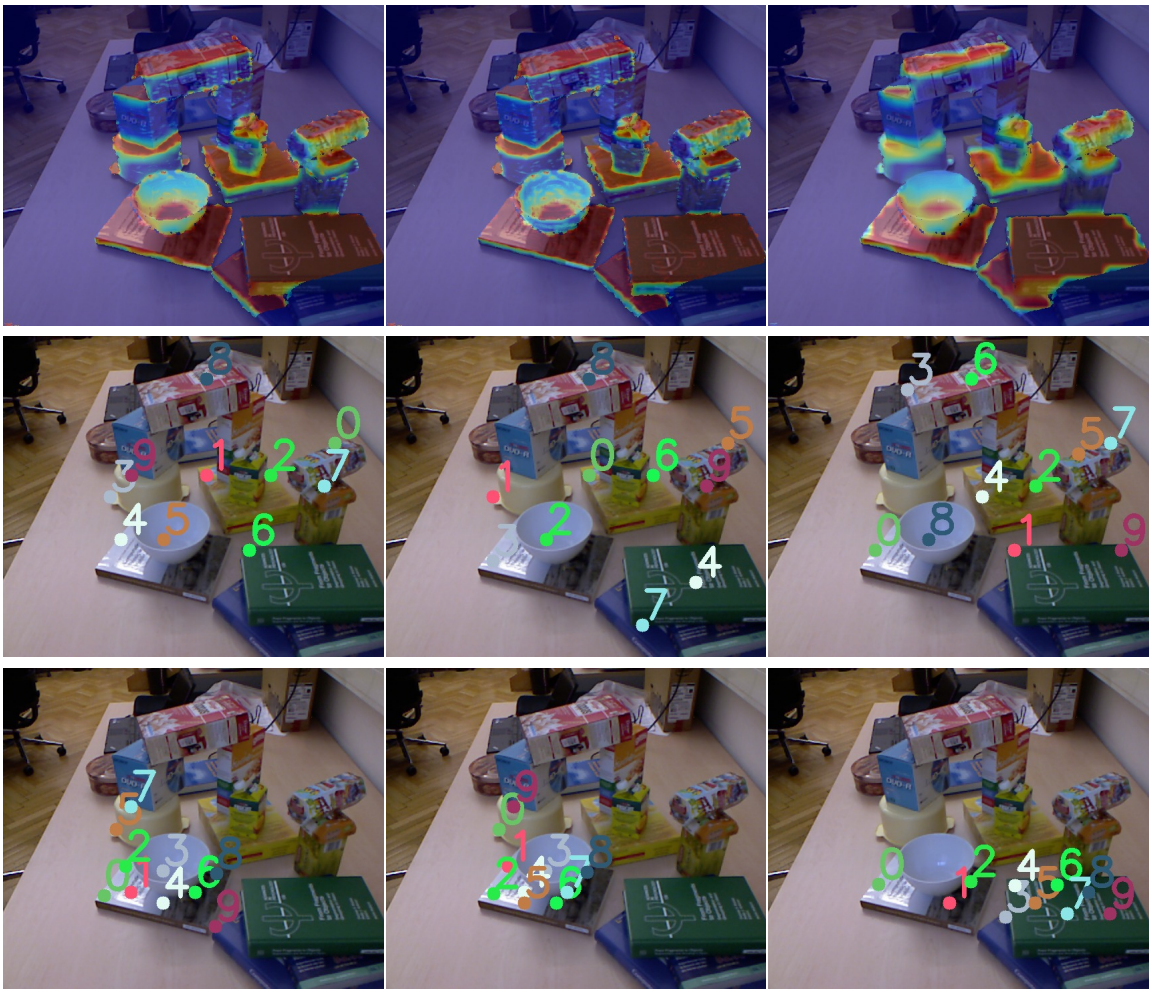


Figure 2.22: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color. (Numbers in brackets represent average values and standard deviation.)



SIMPLE + SUM + LIN SIMPLE + SUM + NMS SIMPLE + SUM + NLM

Figure 2.23: Row 1 shows examples of Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



SIMPLE + MAX + LIN SIMPLE + MAX + NMS SIMPLE + MAX + NLM

Figure 2.24: Row 1 shows examples of Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

Multi-level Relative Surface Orientation Saliency Maps using Itti-based Feature Pyramid

Hit Ration

The evaluation of multi-level Relative Surface Orientation saliency maps (Fig. 2.25) using Itti-based Feature Pyramid (ITTI) shows that the best performance in terms of the Hit Ratio is achieved when conspicuity maps on different levels are combined using across-scale summation (SUM) or maximization (MAX) with non-linear maximization normalization (NLM), and MSR extraction strategy is used. However, if normalization is changed to non-maxima suppression (NMS), then the performance drops to the worst. On average, MSR extraction strategy showed better results that WTA, except when non-maxima suppression normalization (NMS) was used.

Distance to Center

Relative Surface Orientation saliency map calculated using maximization (MAX), non-linear maximization normalization (NLM) and WTA strategy gives the best result for the Distance to Center metric. The worst result is achieved when across-scale summation (SUM) with non-maxima suppression (NMS) and WTA extraction are used. No obvious benefits of using one or the other type of combination, normalization or extraction were found.

Results

In general, MSR extraction strategy shows better performance than WTA extraction strategy. No benefits of using a more complicated scheme that involves ITTI pyramid were reported compared to a simpler scheme that uses SIMPLE pyramid.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.26, Fig. 2.27.

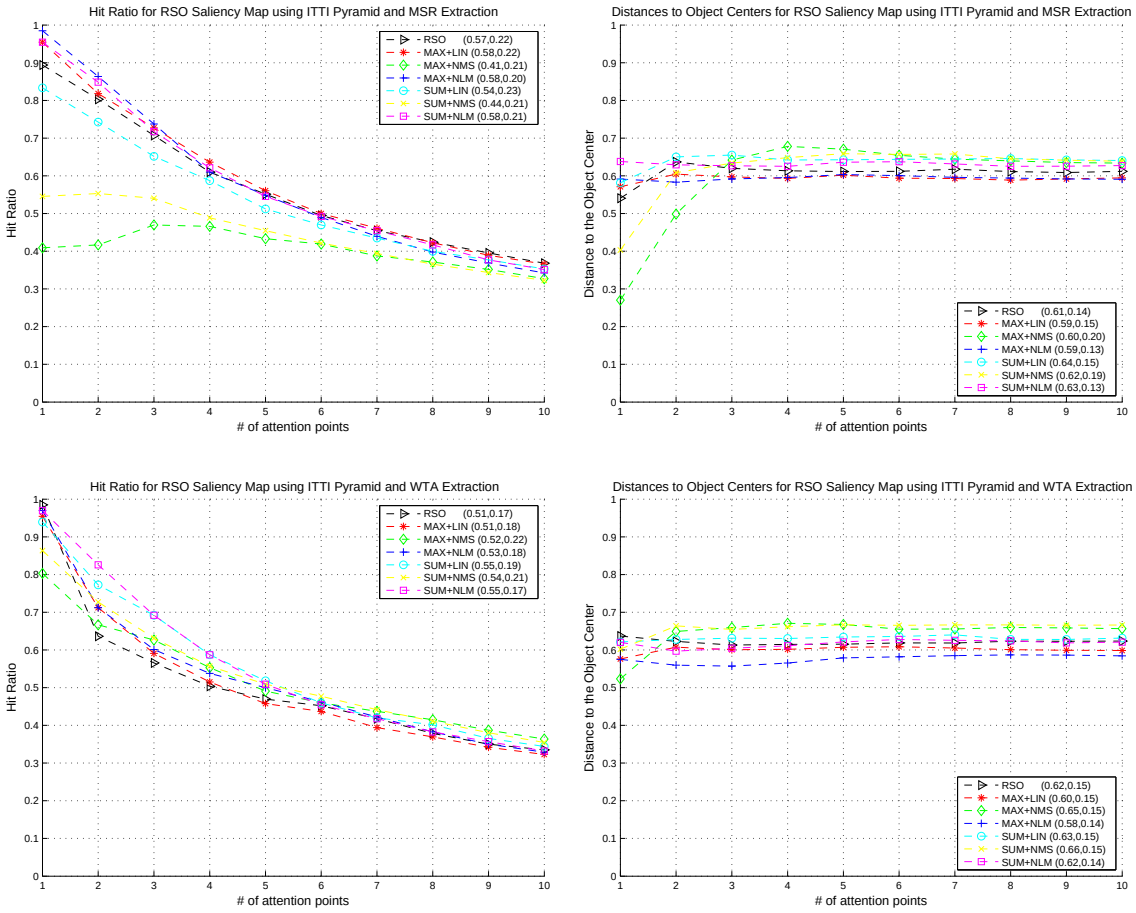
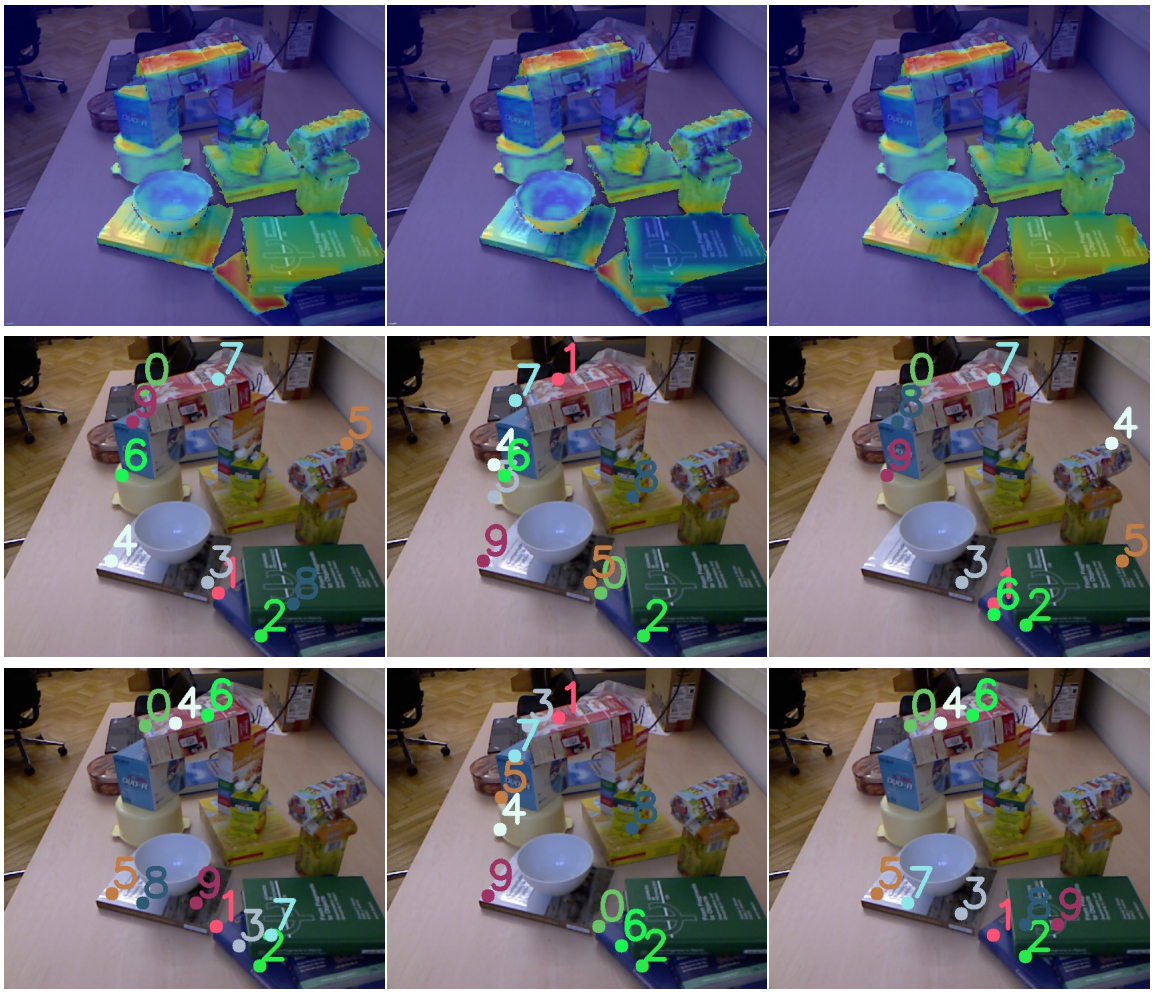
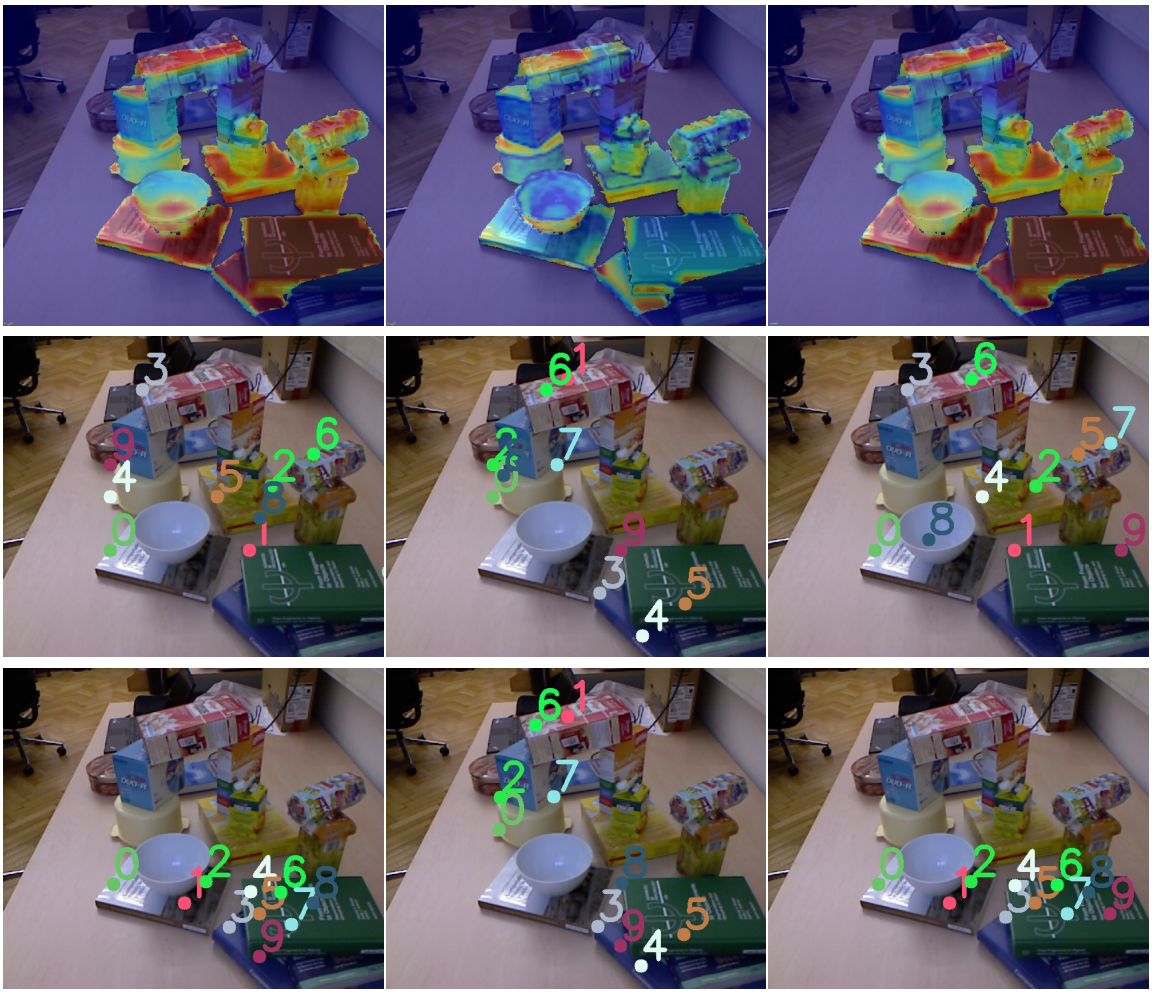


Figure 2.25: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color.



ITTI + SUM + LIN ITTI + SUM + NMS ITTI + SUM + NLM

Figure 2.26: Row 1 shows examples of Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



ITTI + MAX + LIN ITTI + MAX + NMS ITTI + MAX + NLM

Figure 2.27: Row 1 shows examples of Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

2.6.6 3D Symmetry-based Saliency Maps

In this Section we present evaluation results for 3D Symmetry-based saliency maps (SYM3D, Sec. 2.3.4). We used the following multi-level approaches to create feature maps (Sec. 2.4.2): (1) Simple Feature Pyramid $\{F_l\}$ (SIMPLE) and (2) Itti-based Feature Pyramid $\{F_{c,s}\}$ (ITTI).

Combination of levels in the feature pyramid was done using two types of techniques (Sec. 2.4.4): (1) across-scale addition (SUM) and (2) maximization (MAX).

Normalization of saliency maps was done using the following normalization operators (Sec. 2.4.3): (1) linear normalization (LIN), (2) non-maxima suppression (NMS) and (3) non-linear maximization (NLM).

Attention points for object detection were extracted using three different methods: (1) Winner-Take-All (WTA, Sec. 2.5.1), (2) Most Salient Region (MSR, Sec. 2.5.2) and (3) T-junctions (TJ, Sec. 2.5.3).

Our goal in this Section is to understand what multi-level approach, combination and normalization methods along with extraction strategy are the most suitable for 3D Symmetry-based saliency maps when used for object detection. We also want to understand if the usage of more complicated strategies results in increased performance.

Multi-level 3D Symmetry-based Saliency Maps using Simple Feature Pyramid

Hit Ratio

The evaluation of multi-level 3D Symmetry-based saliency maps (Fig. 2.28) using Simple Feature Pyramid (SIMPLE) shows that the best performance in terms of Hit Ratio is achieved when conspicuity maps on different levels are combined using across-scale addition (SUM) with linear normalization (LIN) and TJ extraction strategy. The worst performance in terms of the Hit Ratio is shown when MSR extraction strategy is applied to saliency maps based on combination by maximization (MAX) with non-linear maximization normalization (NLM). For MSR and WTA extraction strategies normalization did not play a significant role for the performance. This can be explained by the nature of the 3D Symmetry-based saliency map, *i. e.* it consists of a number of disjoint components, each having one or several strong local maxima which will be detected by MSR or WTA algorithms. However, normalization plays more important role for TJ extraction strategy, because different normalizations may result in different number of connected components and therefore different TJ points are calculated. In general, when TJ extraction strategy is applied, combination of the pyramid by across-scale addition result in better detection performance, than combination by maximization.

Distance to Center

In terms of the Distance to Center metric, across-scale summation (SUM) with non-linear normalization (NLM) and TJ strategy gives the best result. The worst results are achieved when one-level 3D Symmetry-based saliency map is used either with MSR or WTA extraction strategies. Overall, for a given extraction strategy there are no preferences in selecting

a specific type of pyramid combination and normalization method. The choice of extraction strategy is more important, and TJ strategy clearly outperforms MSR and WTA algorithms.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.29, Fig. 2.30.

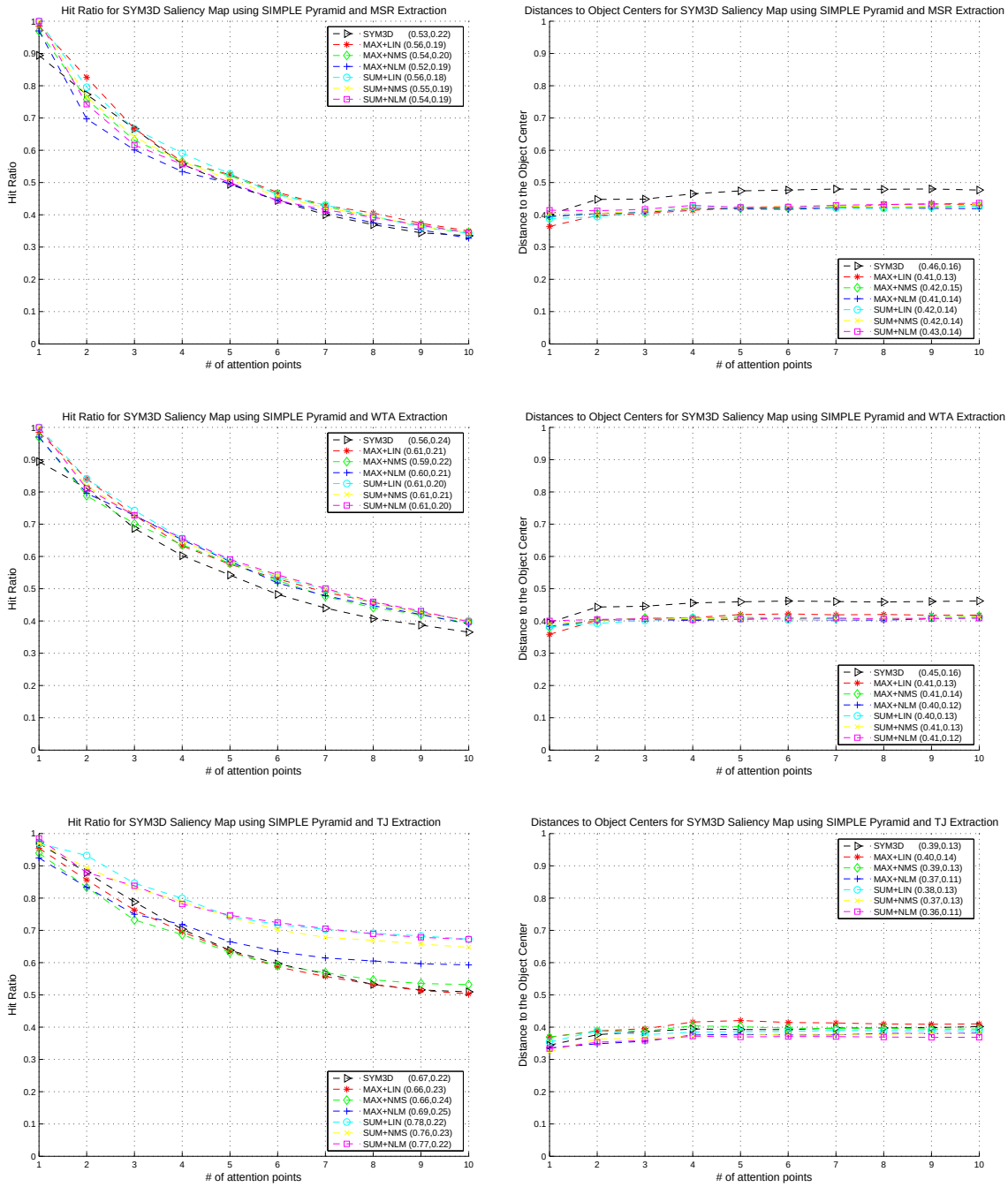
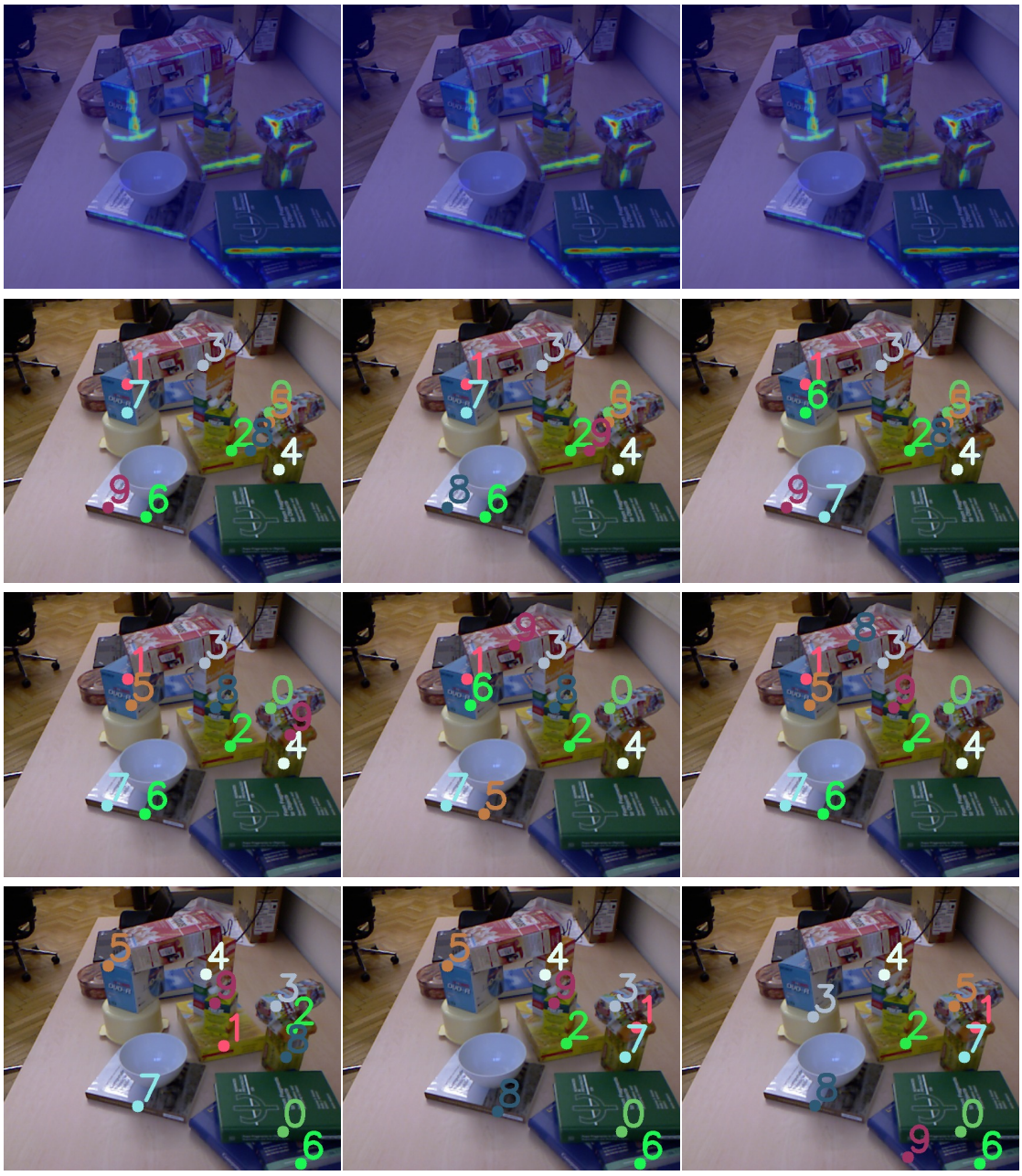
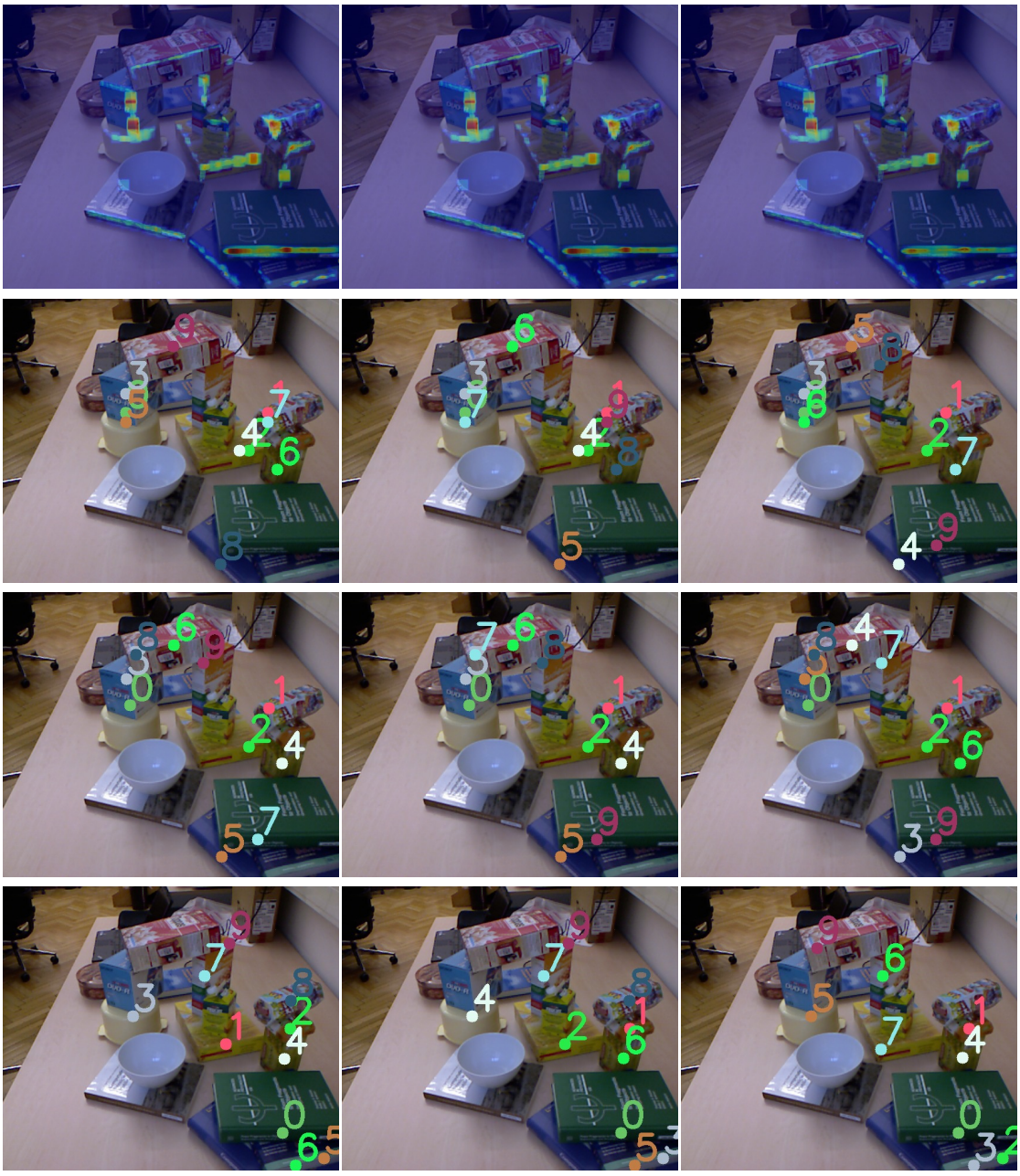


Figure 2.28: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for 3D Symmetry-based saliency map (SYM3D, Sec. 2.3.4) calculated using $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2). Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy, row 2 for WTA (Sec. 2.5.1) and row 3 for TJ (Sec. 2.5.3) extraction strategy. Please note that pictures are best seen in color. (Numbers in brackets represent average values and standard deviation.)



SIMPLE + SUM + LIN SIMPLE + SUM + NMS SIMPLE + SUM + NLM

Figure 2.29: Row 1 shows examples of 3D Symmetry-based saliency map (SYM3D, Sec. 2.3.4) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2, 3 and 4 show first ten attention points extracted using MSR (Sec. 2.5.2), WTA (Sec. 2.5.1) and TJ (Sec. 2.5.3) extraction strategies respectively. Please note that pictures are best seen in color.



SIMPLE + MAX + LIN SIMPLE + MAX + NMS SIMPLE + MAX + NLM

Figure 2.30: Row 1 shows examples of 3D Symmetry-based saliency map (SYM3D, Sec. 2.3.4) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_l\}$ pyramid (SIMPLE, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2, 3 and 4 show first ten attention points extracted using MSR (Sec. 2.5.2), WTA (Sec. 2.5.1) and TJ (Sec. 2.5.3) extraction strategies respectively. Please note that pictures are best seen in color.

Multi-level 3D Symmetry-based Saliency Maps using Itti-based Feature Pyramid

Hit Ratio

The evaluation of multi-level 3D Symmetry-based saliency maps (Fig. 2.31) using Itti-based Feature Pyramid (ITTI) shows that the best performance in terms of the Hit Ratio is achieved when conspicuity maps on different levels are combined using across-scale summation (SUM) with non-linear maximization normalization (NLM) and TJ extraction strategy. When combination by maximization (MAX) is used with non-maxima suppression (NMS) and MSR extraction strategy , the performance drops to the worst. In general, TJ strategy results in up to 100% better performance than other strategies. It is worth mentioning that when TJ strategy is used, the Hit Ratio slightly drops after several first attention points and then maintains at the constant level. This means that there is a small number of repetitive detection of the same object.

Distance to Center

In terms of Distance to Center metric, maximization (MAX) with non-maxima suppression normalization (NMS) and TJ strategy gives the best result. If one-level saliency map with MSR or WTA extraction strategy is used the Distance to the Center metric shows the worst performance. TJ extraction strategy results in overall better precision during the detection.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.32, Fig. 2.33.

Results

Evaluation showed that for 3D Symmetry-based saliency maps TJ extraction strategy significantly improves detection results both in terms of Hit Ratio and Distance to Center metrics. Moreover, for this particular type of saliency maps, the usage of more complicated pyramid type, *i. e.* Itti-based Feature Pyramid (ITTI), resulted in the performance increase up to 10%.

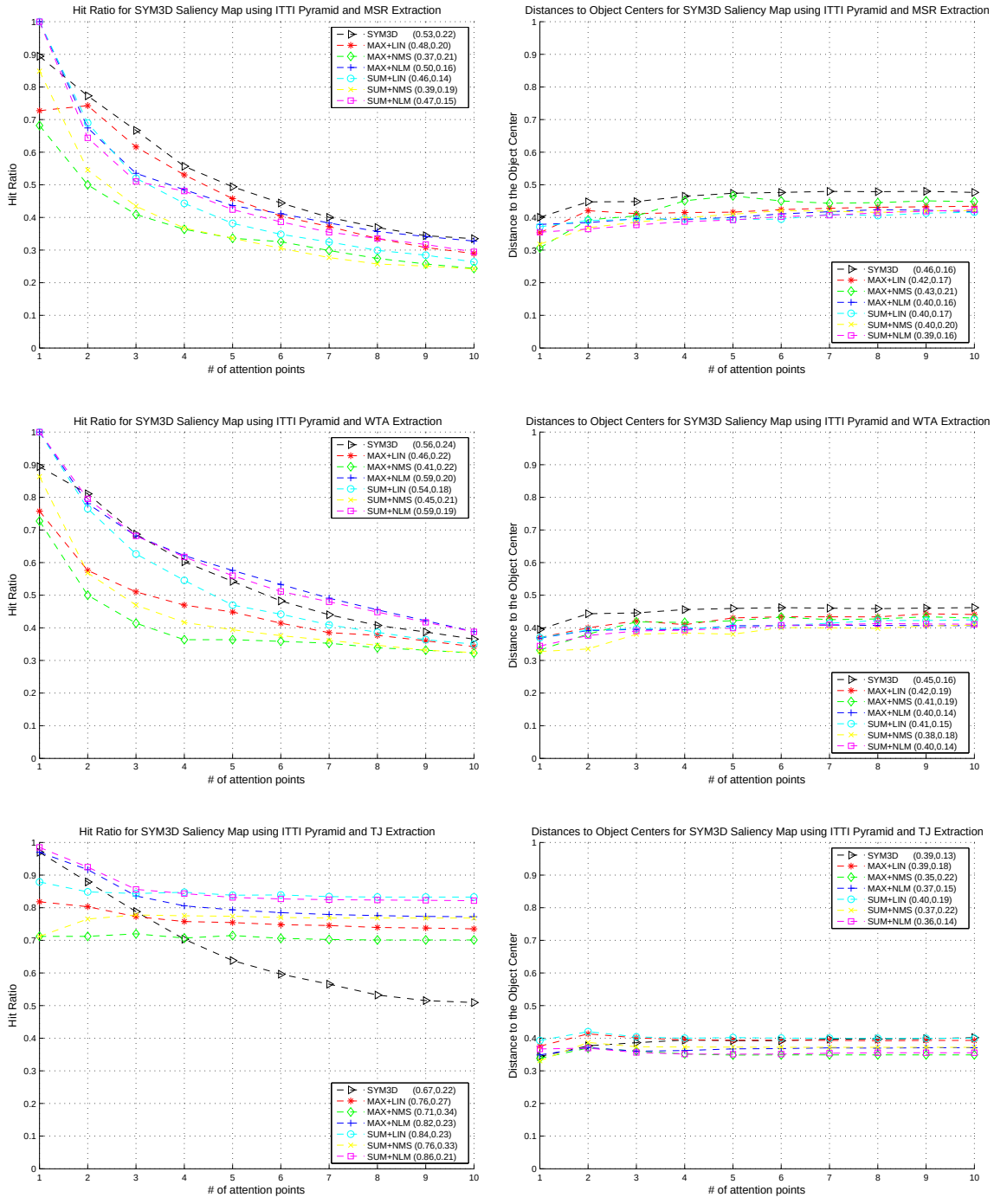
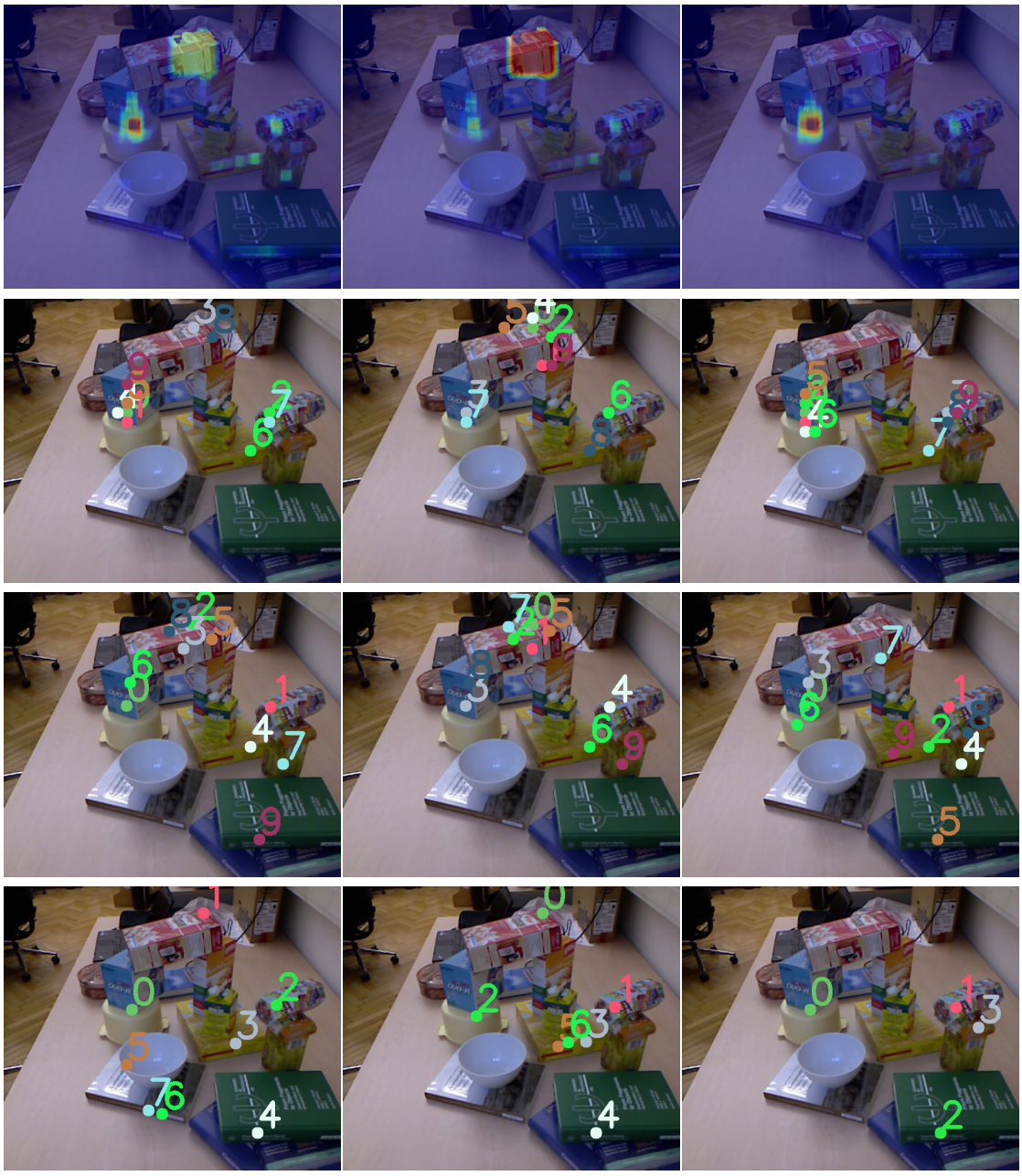
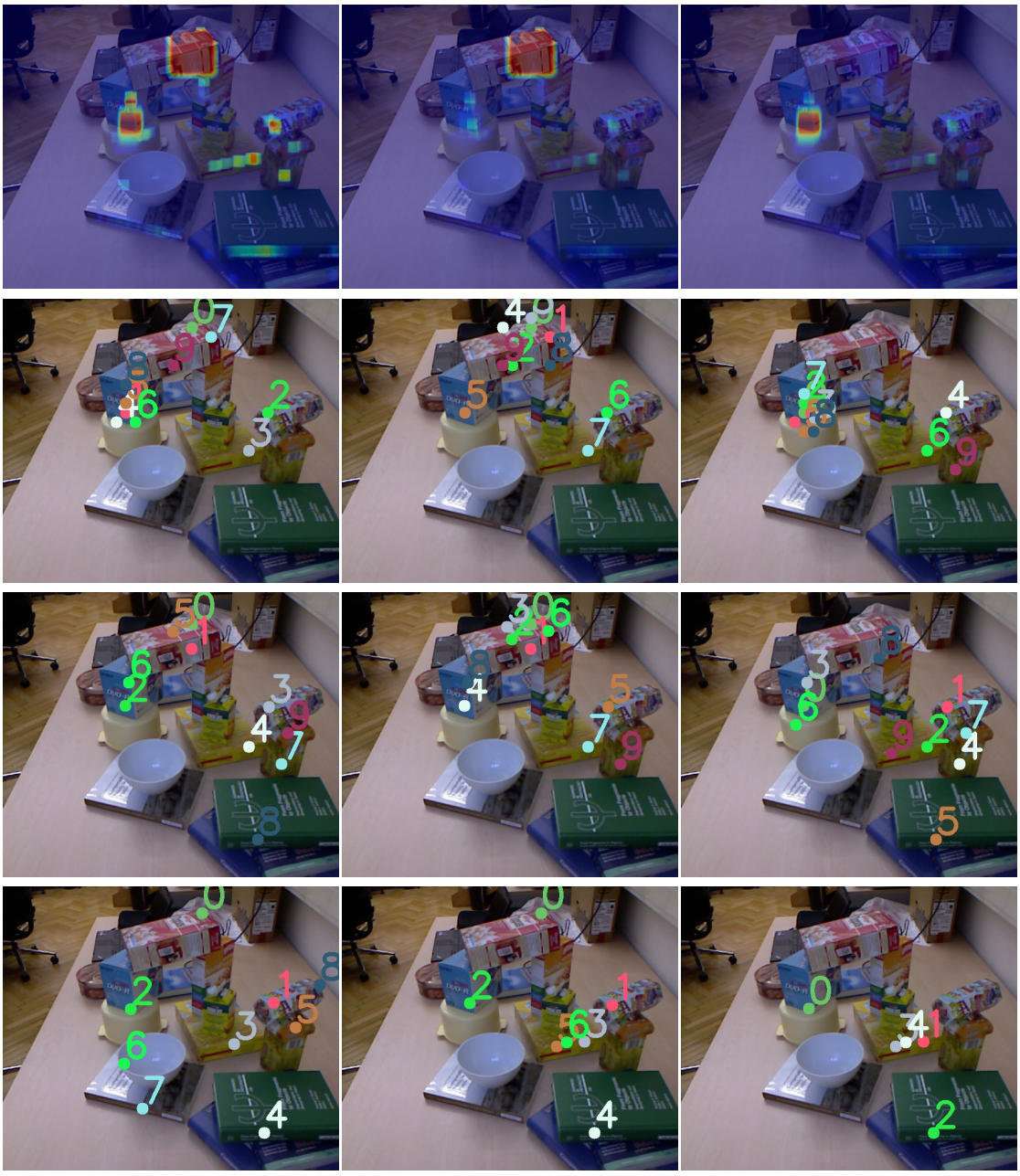


Figure 2.31: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) calculated using $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2). Row 1, row 2 and row 3 show results for MSR (Sec. 2.5.2), WTA (Sec. 2.5.1) and TJ (Sec. 2.5.3) extraction strategies. Please note that pictures are best seen in color.



ITTI + SUM + LIN ITTI + SUM + NMS ITTI + SUM + NLM

Figure 2.32: Row 1 shows examples of 3D Symmetry-based saliency map (SYM3D, Sec. 2.3.4) calculated using across-scale addition (SUM, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2, 3 and 4 show first ten attention points extracted using MSR (Sec. 2.5.2), WTA (Sec. 2.5.1) and TJ (Sec. 2.5.3) extraction strategies respectively. Please note that pictures are best seen in color.



ITTI + MAX + LIN ITTI + MAX + NMS ITTI + MAX + NLM

Figure 2.33: Row 1 shows examples of 3D Symmetry-based saliency map (SYM3D, Sec. 2.3.4) calculated using across-scale addition (MAX, Sec. 2.4.4) of $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2) overlaid with images from the Object Segmentation Database (OSD). Columns 1-3 represent linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear maximization normalization (NLM) respectively (Sec. 2.4.3). Row 2, 3 and 4 show first ten attention points extracted using MSR (Sec. 2.5.2) WTA (Sec. 2.5.1) and TJ (Sec. 2.5.3) extraction strategies respectively. Please note that pictures are best seen in color.

2.6.7 Combinations of Proposed Saliency Maps

In this Section we evaluate the performance of different combinations (Sec. 2.4.5) of the proposed saliency maps. We use the following saliency maps to create a master map:

- *Point-based Height Saliency Map* (PH, Sec. 2.3.1)
- *Surface-based Height Saliency Map* (SH, Sec. 2.3.2)
- *Relative Surface Orientation Saliency Map* (RSO, Sec. 2.3.3)
- Harel *et al.* (HAREL, [Harel 06])

Combination of saliency maps was done using two types of techniques (Sec. 2.4.4): (1) linear combination (SUM) and (2) maximization (MAX).

Normalization of saliency maps was done using the following normalization operators (Sec. 2.4.3): (1) linear normalization (LIN), (2) non-maxima suppression (NMS) and (3) non-linear maximization (NLM).

Attention points for object detection were extracted using two different methods: (1) Winner-Take-All (WTA, Sec. 2.5.1), (2) Most Salient Region (MSR, Sec. 2.5.2).

Our goal in this Section is to understand if a combination of proposed saliency maps increases the performance of object detection. We also want to investigate, if combining color-based saliency maps with 3D-based saliency maps will result in better object detection.

Hit Ratio

The evaluation (Fig. 2.34) shows that the best performance in terms of the Hit Ratio is shown for a master saliency map calculated both using 3D-based and color-based saliency maps, when linear combination with non-linear maximization (NLM) is used together with MSR strategy. The worst performance in terms of the Hit Ratio is shown when MSR extraction strategy is applied to master saliency map calculated only using pure 3D-based saliency maps by means of linear combination (SUM) with linear normalization (LIN). In general, evaluation showed that using color information improves detection results. Moreover, this combination is preferably to be done using linear combinations. No significant differences or trends in performance were discovered between MSR and WTA extraction strategies.

Distance to Center

In terms of the Distance to Center metric the best performance is shown when 3D-based saliency maps are linearly (SUM) combined with color-based saliency map using non-linear maximization (NLM) and attention points are extracted by MSR strategy. Please note that for this type of combination, performance of Hit Ratio was also the best. The worst performance is shown when saliency maps are combined with or without color-based saliency map using maximization (MAX) and non-maxima suppression (NMS) or non-local maximization (NLM) normalizations. Overall, evaluation did not find any significant preferences in combination, normalization and extraction strategies.

Results

Combinations of different types of saliency maps did not show significant improvement compared to single saliency maps. Adding the color-based saliency map slightly improved the object detection performance. Linear combination was found out to be a preferable way to combine maps with different modalities.

Examples of different types of saliency maps and extracted attention points can be seen in Fig. 2.35, Fig. 2.36, Fig. 2.37, Fig. 2.38.

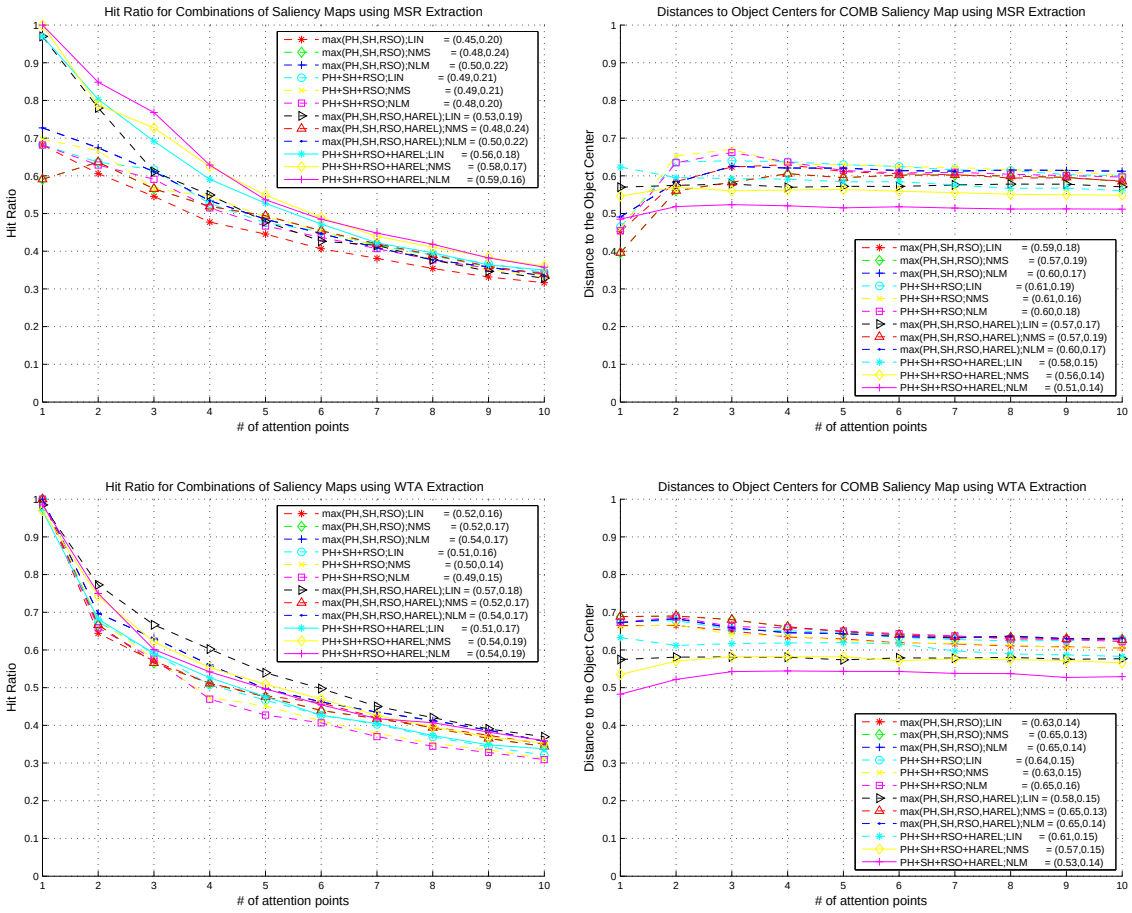
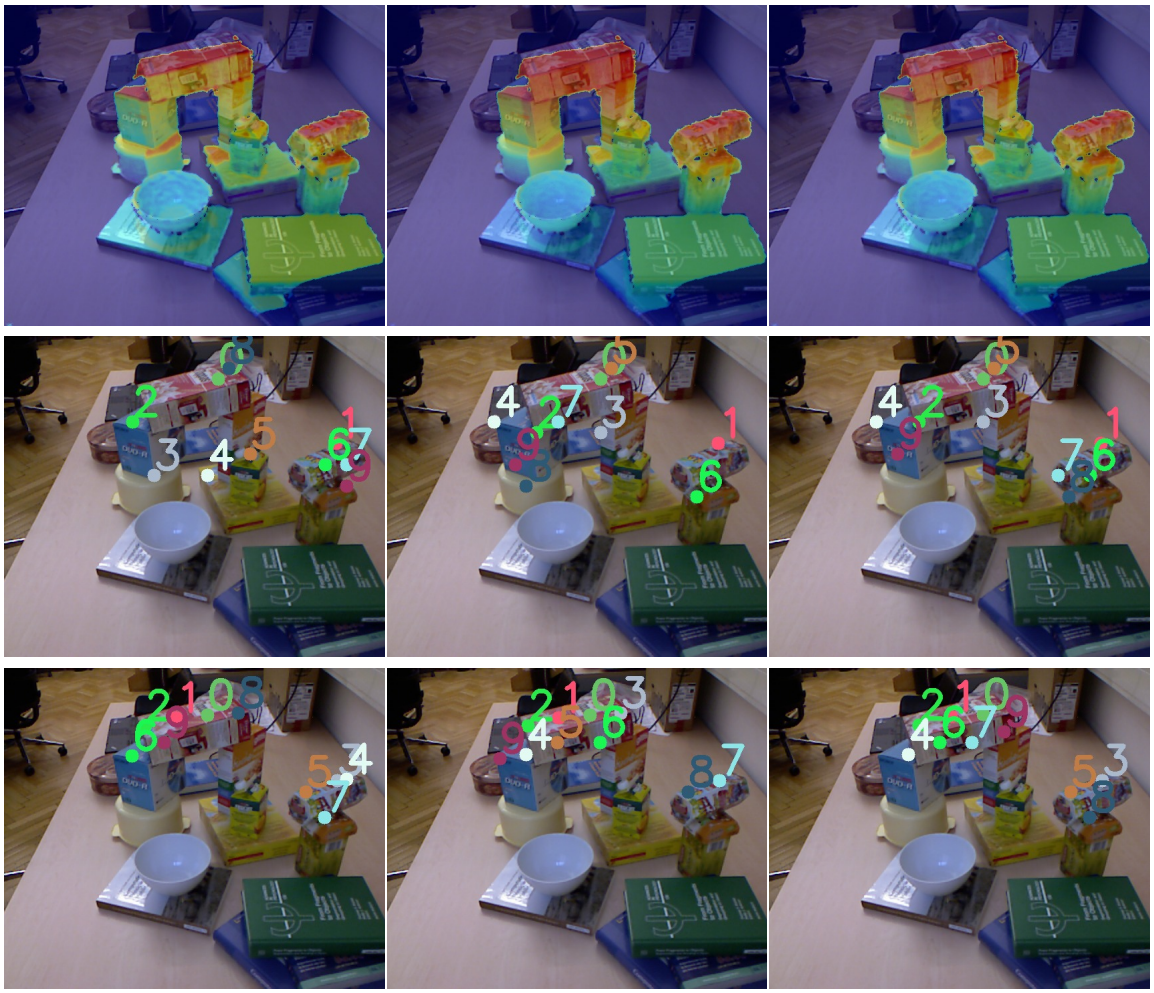
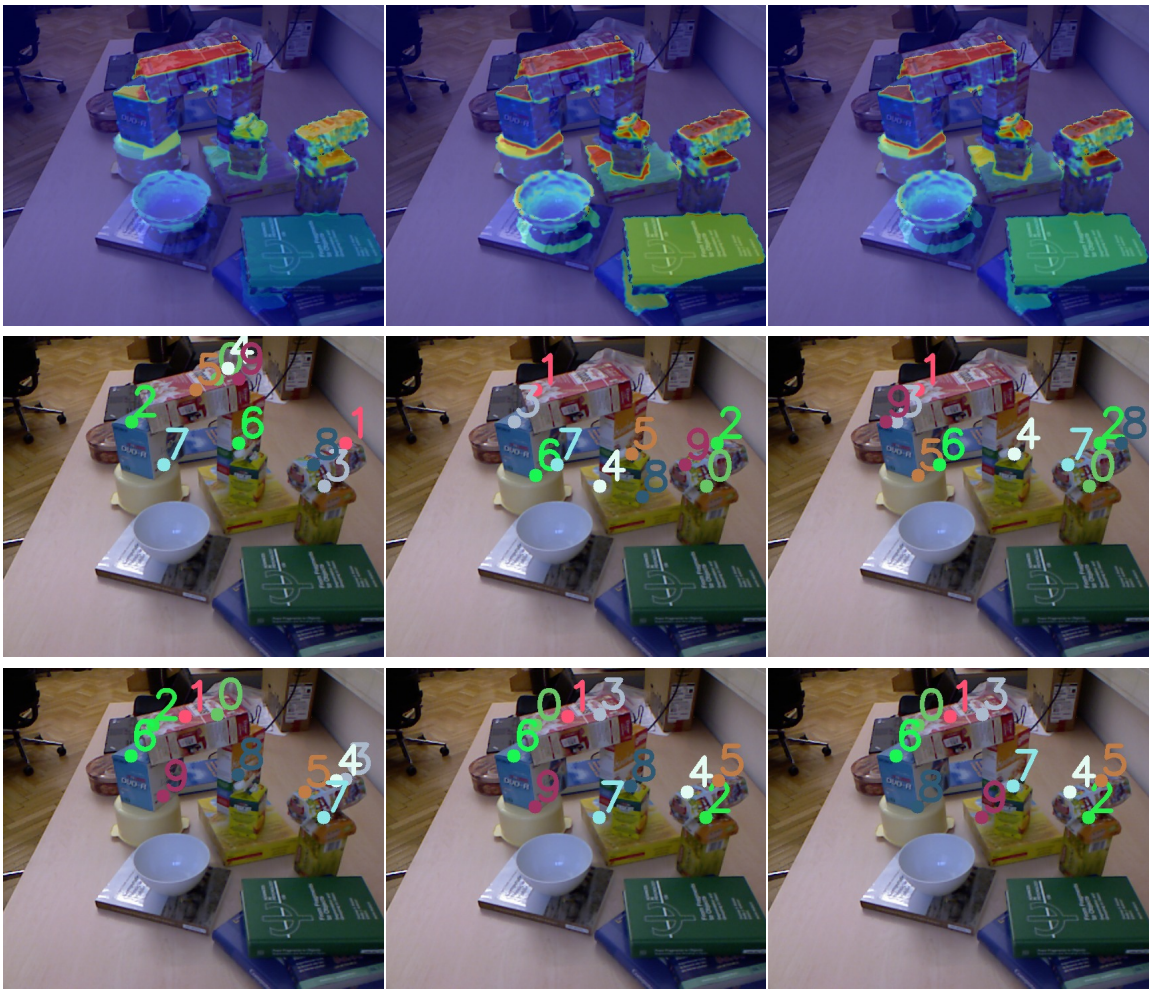


Figure 2.34: Column 1 and column 2 show respectively averaged Hit Ratio (HR) and averaged Distance to the Center (DC) against the number of extracted attention points for combined saliency map (Sec. 2.4.5) with and without color saliency map proposed by Harel *et al.* [Harel 06] respectively. Row 1 shows results for MSR (Sec. 2.5.2) extraction strategy and row 2 for WTA (Sec. 2.5.1) extraction strategy. Please note that pictures are best seen in color. (Numbers in brackets represent average values and standard deviation.)



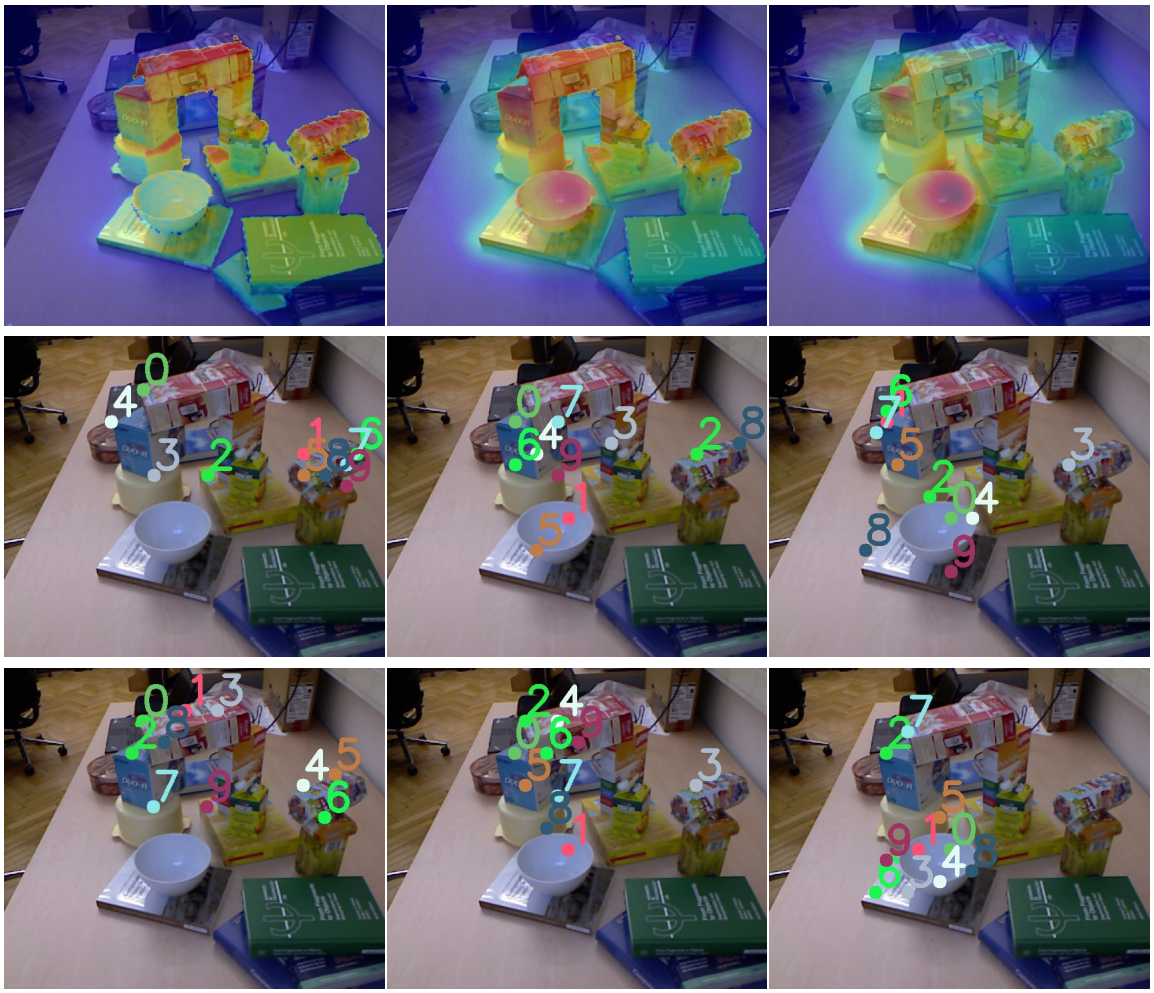
SUM + LIN SUM + NMS SUM + NLM

Figure 2.35: Row 1 shows examples of combined saliency map calculated using linear combination (COMB, SUM, Sec. 2.4.5) overlaid with images from the Object Segmentation Database (OSD). Column 1, 2 and 3 show linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



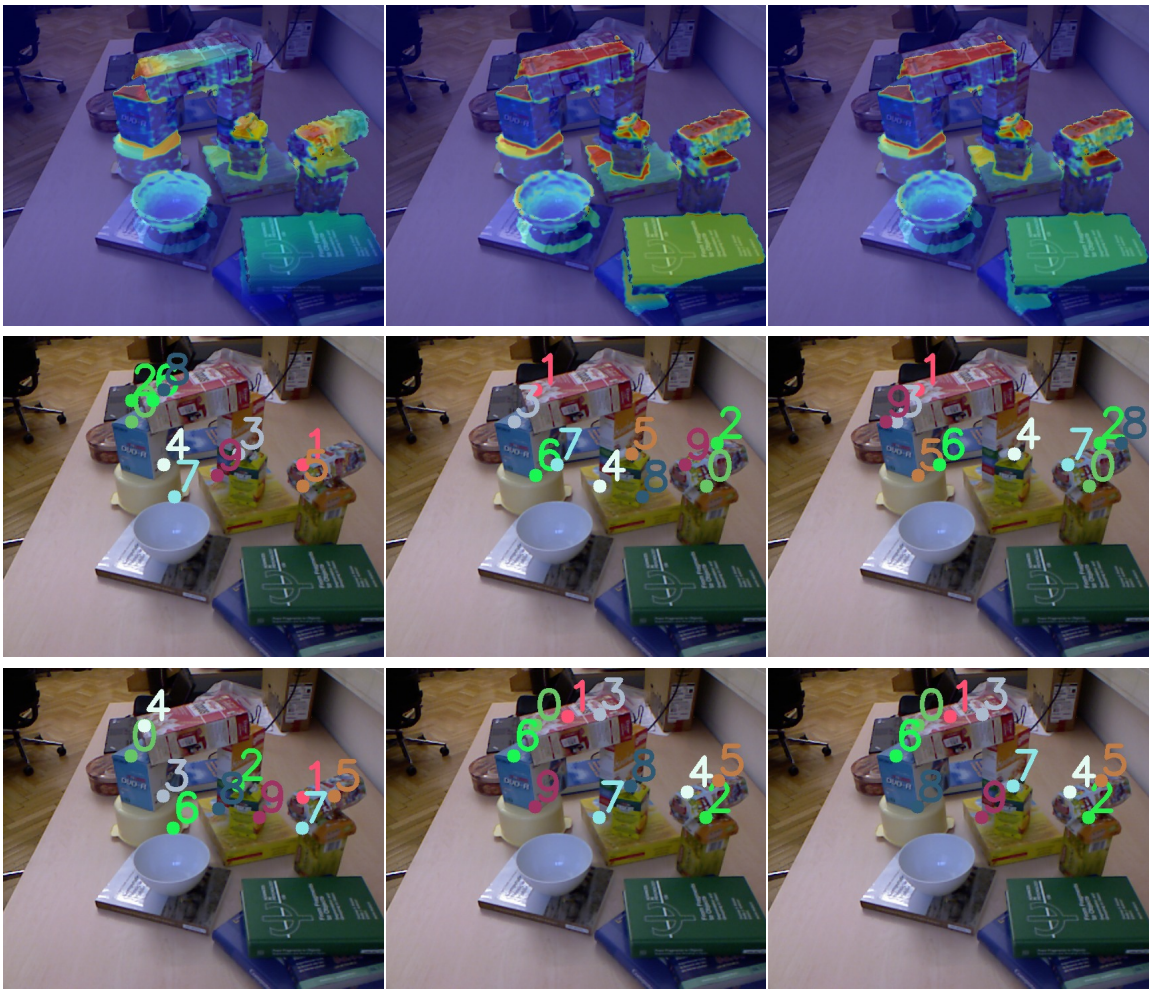
MAX + LIN MAX + NMS MAX + NLM

Figure 2.36: Row 1 shows examples of combined saliency map calculated using maximization (COMB, MAX, Sec. 2.4.5) overlaid with images from the Object Segmentation Database (OSD). Column 1, 2 and 3 show linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



SUM + HAREL + LIN SUM + HAREL + NMS SUM + HAREL + NLM

Figure 2.37: Row 1 shows examples of combined saliency map including color saliency map proposed by Harel *et al.* [Harel 06] calculated using linear combination (COMB+HAREL, SUM, Sec. 2.4.5) overlaid with images from the Object Segmentation Database (OSD). Column 1, 2 and 3 show linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.



MAX + HAREL + LIN MAX + HAREL + NMS MAX + HAREL + NLM

Figure 2.38: Row 1 shows examples of combined saliency map including color saliency map proposed by Harel *et al.* [Harel 06] calculated using maximization (COMB+HAREL, MAX, Sec. 2.4.5) overlaid with images from the Object Segmentation Database (OSD). Column 1, 2 and 3 show linear normalization (LIN), non-maxima suppression normalization (NMS) and non-linear normalization (NLM) respectively (Sec. 2.4.3). Row 2 and 3 show first ten attention points extracted using MSR (Sec. 2.5.2) and WTA (Sec. 2.5.1) extraction strategies respectively. Please note that pictures are best seen in color.

2.7 Discussion

We investigated the use of 3D cues to obtain attention points that can be used further for object detection and as seed points for object segmentation. New attention algorithms based on 3D data were proposed: Point-based Height saliency map (PH, Sec. 2.3.1), Surface-based Height saliency map (SH, Sec. 2.3.2), Relative Surface Orientation saliency map (RSO, Sec. 2.3.3) and 3D Symmetry-based saliency (SYM3D, Sec. 2.3.4). We investigated two different strategies to build multi-level saliency maps (Sec. 2.4.2): Simple Feature Pyramid (SIMPLE) and Itti-based Feature Pyramid (ITTI). Pyramid levels were combined using two different approaches (Sec. 2.4.4): across-scale addition (SUM) and maximization (MAX). Three different normalizations methods were used (Sec. 2.4.3): linear normalization (LIN), non-maxima suppression (NMS) and non-linear maximum (NLM). Combinations of PH, SH and RSO saliency maps into a master saliency map (Sec. 2.4.5) were investigated with and without 2D-based state-of-the-art saliency map (HAREL [Harel 06]).

To extract attention points from saliency maps we used three different types of strategies: Winner-Take-All (WTA, 2.5.1), Most Salient Region (MSR, 2.5.2) and a new proposed strategy based on T-junctions (TJ, Sec. 2.5.3). The quality of attention points was defined in terms of two metrics (Sec. 2.6.1): Hit Ratio (HR) and Distance to Center (DC).

Evaluation was carried out on the *Object Segmentation Database (OSD)*. We compared our approaches to the following state-of-the-art saliency algorithms: Itti *et al.* [Itti 98], Harel *et al.* [Harel 06], Hou *et al.* [Hou 07], Bruce *et al.* [Bruce 09], Kootstra *et al.* [Kootstra 10b], Wang *et. al* [Wang 13].

As can be seen from the Tables 2.4 and 2.5 the use of 3D-based saliency maps improves object detection performance compared to color-based state-of-the-art saliency maps. The best Hit Ratio (54%) for color-based saliency operators was shown by Harel attention model [Harel 06], while 3D Symmetry-based saliency map together with the extraction strategy based on T-Junctions resulted in a Hit Ratio of 86%. This basically means that the number of correctly detected objects increased by 60%. Detection accuracy represented by Distance to Center metric was improved by 10% when 3D Symmetry-based saliency map was used, meaning that attention points are situated closer to object centers. These results show that 3D symmetry-based saliency maps capture the main structure of table scenes with man-made objects and direct attention to those objects.

Furthermore, we can conclude from plots (Fig. 2.5, Fig. 2.9, Fig. 2.13, Fig. 2.16, Fig. 2.19, Fig. 2.22, Fig. 2.25, Fig. 2.28, Fig. 2.31, Fig. 2.34) that the WTA extraction strategy does not show a significantly better performance, than the MSR extraction strategy. No significant performance benefits were achieved when the ITTI pyramid was used instead of the SIMPLE pyramid. With respect to normalization, we conclude that linear normalization (LIN) shows the same results as non-maxima suppression (NMS) or non-linear maximization (NLM). This means that the choice of normalization does not change results significantly. Overall, for the proposed 3D-based saliency maps, using the SIMPLE pyramid with linear normalization (LIN) is sufficient.

With respect to the Distance to Center metric, all detection methods, except for those that used 3D Symmetry-based saliency map, resulted in similar performance. For the 3D Symmetry-based saliency map all extraction strategies showed improved performance.

This effect can be explained by the nature of 3D Symmetry-based saliency in which salient regions represent symmetry lines of objects. All described attention points extraction strategies detect points on these symmetry lines. Therefore, attention points are located closer to the object center.

No significant improvements were shown for the combination of the proposed saliency maps with color-based saliency maps.

With all of the given results, we can conclude it is not sufficient to use only color-based saliency maps to detect objects. Adding 3D information for saliency computation increases the detection performance. We found that multi-level strategy and different types of combination and normalization do not play an important role. It was also shown that there is no benefit in using more sophisticated, and therefore more computationally expensive extraction strategies.

The proposed method to detect objects using 3D Symmetry-based saliency maps and T-Junction extraction strategy clearly outperforms all other methods. It was specifically designed to detect objects in cluttered table scenes containing man-made objects. This means that saliency maps involving 3D information should be developed and used to detect objects in robotic tasks. Those saliency maps should specifically capture object properties, instead of single features, to guide attention to objects.

After we detect objects we want to manipulate them. For this we need their shapes or outlines. This problem is solved by segmenting them from the scene using attention-driven segmentation.

SALIENCY MAP			NORMALIZATION											
			–			LIN			NMS			NLM		
			WTA	MSR	TJ	WTA	MSR	TJ	WTA	MSR	TJ	WTA	MSR	TJ
2D&3D STATE-OF-THE-ART SALIENCY		ITTI [ITTI 98]	**0.49**	0.41	–	–	–	–	–	–	–	–	–	–
		HAREL [HAREL 06]	**0.54**	**0.54**	–	–	–	–	–	–	–	–	–	–
		HOU [HOU 07]	**0.36**	0.32	–	–	–	–	–	–	–	–	–	–
		BRUCE [BRUCE 09]	**0.48**	0.42	–	–	–	–	–	–	–	–	–	–
		KOOTSRA [KOOTSTRA 10B]	0.51	**0.53**	–	–	–	–	–	–	–	–	–	–
		WANG [WANG 13]	0.48	**0.50**	–	–	–	–	–	–	–	–	–	–
		WANG+ITTI [WANG 13]	**0.53**	0.50	–	–	–	–	–	–	–	–	–	–
		WANG+HOU [WANG 13]	**0.53**	0.50	–	–	–	–	–	–	–	–	–	–
		WANG+BRUCE[WANG 13]	**0.54**	0.51	–	–	–	–	–	–	–	–	–	–
SINGLE MAP		PH	**0.45**	0.44	–	–	–	–	–	–	–	–	–	–
		SH	0.40	**0.49**	–	–	–	–	–	–	–	–	–	–
		RSO	0.51	**0.57**	–	–	–	–	–	–	–	–	–	–
		SYM3D	0.56	0.53	**0.67**	–	–	–	–	–	–	–	–	–
SIMPLE PYRAMID TYPE	SUM	PH	–	–	–	0.46	**0.58**	–	0.45	0.57	–	0.46	0.55	–
		SH	–	–	–	0.41	**0.56**	–	0.40	0.55	–	0.41	0.54	–
		RSO	–	–	–	0.52	**0.62**	–	0.51	0.61	–	0.52	0.61	–
		SYM3D	–	–	–	0.61	0.56	**0.78**	0.61	0.55	0.76	0.61	0.54	0.77
	MAX	PH	–	–	–	0.45	0.41	–	0.44	0.43	–	0.46	**0.55**	–
		SH	–	–	–	0.39	0.46	–	0.39	0.48	–	0.41	**0.51**	–
		RSO	–	–	–	0.48	0.52	–	0.49	0.58	–	0.52	**0.60**	–
		SYM3D	–	–	–	0.61	0.56	0.66	0.59	0.54	0.66	0.60	0.52	**0.69**
ITTI PYRAMID TYPE	SUM	PH	–	–	–	0.52	0.44	–	**0.53**	0.35	–	0.51	0.49	–
		SH	–	–	–	**0.47**	0.43	–	**0.47**	0.41	–	0.45	**0.47**	–
		RSO	–	–	–	0.55	0.54	–	0.54	0.44	–	0.55	**0.58**	–
		SYM3D	–	–	–	0.54	0.46	0.84	0.45	0.39	0.76	0.59	0.47	**0.86**
	MAX	PH	–	–	–	0.45	**0.56**	–	0.52	0.33	–	0.46	0.55	–
		SH	–	–	–	0.43	**0.50**	–	0.47	0.39	–	0.43	0.48	–
		RSO	–	–	–	0.51	**0.58**	–	0.52	0.41	–	0.53	**0.58**	–
		SYM3D	–	–	–	0.46	0.48	0.76	0.41	0.37	0.71	0.59	0.50	**0.82**
COMBINATIONS	SUM	–	–	–	–	**0.51**	0.49	–	0.50	0.49	–	0.49	0.48	–
		+HAREL	–	–	–	0.52	0.56	–	0.54	0.58	–	0.54	**0.59**	–
	MAX	–	–	–	–	0.51	0.45	–	0.52	0.48	–	**0.54**	0.50	–
		+HAREL	–	–	–	**0.57**	0.53	–	0.52	0.48	–	0.54	0.50	–

Table 2.4: The table shows comparative results for Hit Ratio for all evaluated methods. As can be seen from the table, 3D Symmetry-based Saliency map in combination with $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2), across-scale summation (SUM, Sec. 2.4.4) and non-linear maximization normalization (NLM, Sec. 2.4.3) gives up to 100% better performance than other methods when TJ extraction strategy (Sec. 2.5.3) is applied. Bold numbers represent the best Hit Ratio for each type of saliency map within a specific combination type across different normalization and extraction strategies. The best and the worst Hit Ratios for each type of saliency map regardless of combination type, normalization and extraction strategies are highlighted in green and red respectively.

SALIENCY MAP			NORMALIZATION											
			–			LIN			NMS			NLM		
			WTA	MSR	TJ	WTA	MSR	TJ	WTA	MSR	TJ	WTA	MSR	TJ
2D&3D STATE-OF-THE-ART SALIENCY		ITTI [ITTI 98]	0.48	**0.47**	–	–	–	–	–	–	–	–	–	–
		HAREL [HAREL 06]	**0.52**	0.53	–	–	–	–	–	–	–	–	–	–
		HOU [HOU 07]	0.47	**0.46**	–	–	–	–	–	–	–	–	–	–
		BRUCE [BRUCE 09]	**0.47**	0.48	–	–	–	–	–	–	–	–	–	–
		KOOTSRA [KOOTSTRA 10B]	0.41	**0.40**	–	–	–	–	–	–	–	–	–	–
		WANG [WANG 13]	0.42	**0.41**	–	–	–	–	–	–	–	–	–	–
		WANG+ITTI [WANG 13]	**0.47**	**0.47**	–	–	–	–	–	–	–	–	–	–
		WANG+HOU [WANG 13]	0.50	**0.49**	–	–	–	–	–	–	–	–	–	–
		WANG+BRUCE[WANG 13]	**0.44**	0.46	–	–	–	–	–	–	–	–	–	–
SINGLE MAP		PH	0.70	**0.53**	–	–	–	–	–	–	–	–	–	–
		SH	0.63	**0.61**	–	–	–	–	–	–	–	–	–	–
		RSO	0.62	**0.61**	–	–	–	–	–	–	–	–	–	–
		SYM3D	0.45	0.46	**0.39**	–	–	–	–	–	–	–	–	–
SIMPLE PYRAMID TYPE	SUM	PH	–	–	–	0.66	**0.62**	–	0.68	0.63	–	0.65	0.63	–
		SH	–	–	–	0.63	**0.60**	–	0.63	**0.60**	–	0.62	0.61	–
		RSO	–	–	–	**0.57**	0.58	–	**0.57**	0.59	–	0.58	0.59	–
		SYM3D	–	–	–	0.40	0.42	0.40	0.41	0.42	0.37	0.41	0.43	**0.36**
	MAX	PH	–	–	–	0.71	**0.48**	–	0.71	0.52	–	0.65	0.64	–
		SH	–	–	–	0.66	0.59	–	0.65	**0.58**	–	0.64	0.62	–
		RSO	–	–	–	0.66	0.63	–	0.61	0.61	–	**0.59**	**0.59**	–
		SYM3D	–	–	–	0.41	0.41	0.39	0.41	0.42	**0.35**	0.40	0.41	0.37
ITTI PYRAMID TYPE	SUM	PH	–	–	–	0.68	0.65	–	0.71	**0.63**	–	0.65	0.66	–
		SH	–	–	–	0.66	0.66	–	0.69	0.66	–	**0.63**	**0.63**	–
		RSO	–	–	–	0.63	0.64	–	0.66	**0.62**	–	**0.62**	0.63	–
		SYM3D	–	–	–	0.41	0.40	0.40	0.38	0.40	0.37	0.40	0.39	**0.36**
	MAX	PH	–	–	–	0.66	0.62	–	0.70	**0.61**	–	0.64	0.63	–
		SH	–	–	–	0.65	0.61	–	0.66	0.64	–	0.61	**0.60**	–
		RSO	–	–	–	0.60	0.59	–	0.65	0.60	–	**0.58**	0.59	–
		SYM3D	–	–	–	0.42	0.42	0.39	0.41	0.43	**0.35**	0.40	0.40	0.37
COMBINATIONS	SUM	–	–	–	–	0.64	0.61	–	0.63	0.61	–	0.65	**0.60**	–
		+HAREL	–	–	–	0.61	0.58	–	0.57	0.56	–	0.53	**0.51**	–
	MAX	–	–	–	–	0.63	0.59	–	0.65	**0.57**	–	0.65	0.60	–
		+HAREL	–	–	–	0.58	**0.57**	–	0.65	**0.57**	–	0.65	0.60	–

Table 2.5: The table shows comparative results for the Distance to Center metric for all evaluated methods. The best and the worst performances for each method are highlighted in green and red respectively. As can be seen from the table, 3D Symmetry-based Saliency map in combination with $\{F_{c,s}\}$ pyramid (ITTI, Sec. 2.4.2), maximization (MAX, Sec. 2.4.4) and non-maximum suppression normalization (NMS, Sec. 2.4.3) gives up to 40% better performance than other methods when TJ extraction strategy (Sec. 2.5.3) is applied. Bold numbers represent the best Distance to Center for each type of saliency map within a specific combination type across different normalization and extraction strategies. The best and the worst Distance to Center for each type of saliency map regardless of combination type, normalization and extraction strategies are highlighted in green and red respectively.

Chapter 3

Attention-driven Segmentations

Segmentation of objects from a static scene is a crucial step in many computer vision tasks, such as learning object models, or identifying objects in a scene. Segmentation in cluttered environments is particularly challenging, and yet it is one of the most practically useful tasks for robotics, with the need to find relevant objects quickly amongst a possibly large number of distractors [Katz 13].

Different approaches have been proposed to tackle the object segmentation problem, which can be broadly classified into two groups: *discriminative* and *agglomerative*. Discriminative segmentation algorithms tend to segment the complete scene at once and assign a label to every pixel [Felzenszwalb 04, Shi 00, Arbelaez 12, Comaniciu 02, Malik 01, Triebel 10, Ückermann 12, Richtsfeld 12]. The computational performance of such algorithms tends to drop with increasing scene complexity. Agglomerative segmentation algorithms grow regions from a seed point to segment the foreground object and agglomerative segmentation can be generally divided into two sub-classes: interactive segmentation and active segmentation.

Boykov *et al.* [Boykov 01] and Rother *et al.* [Rother 04] described interactive approaches for segmentation of color images that however require input from the user, which is often not feasible in robotic applications.

Active segmentation, or attention-driven segmentation, is inspired by models of human visual attention, where the scene is first parsed to detect saliency regions and the segmentation is then applied first to the most salient ones [Ko 06, Ouerhani 01, Mishra 09, Mishra 11, Kootstra 10a]. Originally, active segmentation was proposed to tackle scene complexity (Aloimonos *et al.* [Aloimonos 88]) by processing only relevant parts of the scene. Moreover, in robotic applications it is often more efficient to segment precisely the first most important object and leave the rest as a background.

In this thesis, we present three novel methods for attention-driven segmentation for cluttered table scenes: (1) attention-driven segmentation with object detection based on attention points from 3D symmetry saliency maps [Potapova 12b], (2) attention-driven segmentation based on probabilistic edges, and (3) incremental attention-driven segmentation based on complete-scene segmentation proposed by Richtsfeld *et al.* [Richtsfeld 12].

In this chapter the focus is on developing an attention-driven segmentation approach that can be successfully used to segment objects in complex scenes. The chapter is orga-

nized as follows: In Section 3.1, we review related work. Sections 3.2 and 3.3 describe proposed algorithms as well as evaluation in detail.

3.1 Related Work

The focus of this thesis is segmentation of indoor table scenes typical for robotic task environments. Therefore, we primarily concentrate on the work developed for RGB-D data. As was mentioned in the introduction, segmentation algorithms can be broadly classified into two groups: (*discriminative*) those that segment the complete scene at once into multiple objects and (*agglomerative*) attention-driven methods that segment only one or several chosen objects. Complete scene segmentation algorithms need to go through all segmented objects in order to detect their properties and decide on the object of interest. In contrast, attention-driven segmentation focuses on one object at a time and can use its properties to efficiently detect and segment it in the scene. A number of discriminative segmentation algorithms use depth information to boost segmentation performance for complex indoor scenarios. Such algorithms include those proposed in [Strom 10, Triebel 10, Collet 11, Wan 12, Ückermann 12, Richtsfeld 12].

Strom *et al.* [Strom 10] extended the graph-based segmentation algorithm of Felzenszwalb *et al.* [Felzenszwalb 04] by using similarity of point normals additionally to color similarity between points to extract uniform segments. The method however is not suited for scenes containing textured objects. Triebel *et al.* [Triebel 10] described an unsupervised technique to segment and detect objects in indoor environments. First the scene is clustered both in feature and geometric spaces, a Conditional Random Field is used then to assign labels to segmented clusters. Collet *et al.* [Collet 11] introduced a framework to perform scene segmentation into physically meaningful parts, based on appearance and 3D shape. Wan *et al.* [Wan 12] proposed to over-segment an image into superpixels based on mixed intensity and depth values and then merge them into a set of planes according to their co-planarity. The algorithm was designed to segment the room into specific planes and will require a different set of thresholds for any new type of the data. Ueckermann *et al.* [Ückermann 12] proposed to detect edges using surface normals and employed region growing to find regions of smooth surfaces, separated by the detected edges. An adjacency-graph is built from the detected surfaces and heuristic rules are used to split the graph and segment objects. The approach of Richtsfeld *et al.* [Richtsfeld 12] segments the complete scene into parametrised surface models, which are then grouped together using Gestalt principles. Those parametrised models can be used further for object reconstruction and learning of new object models.

A number of segmentation algorithms [Campbell 07, Lai 12a, Karpathy 13, García 13] use multi-view scene representation to segment objects. Campbell *et al.* [Campbell 07] introduced a volumetric graph-based algorithm, where a camera is fixated on the object of interest during a sequence of movements and a color model of the object is learned. Lai *et al.* [Lai 12a] utilized a sliding window detector trained from object views to label objects in the reconstructed 3D scene. Karpathy *et al.* [Karpathy 13] proposed to "discover objects" by decomposing a scene into candidate segments. Each segment is ranked ac-

cording to its "objectness" measure based on compactness, symmetry, smoothness, local and global convexity. The algorithms requires a scene mesh which authors obtained using Kinect Fusion [Newcombe 11]. Garcia *et al.* [García 13] proposed a framework that detects unknown proto-objects in a 3D scene model built using Kinect Fusion [Newcombe 11]. It was shown that the framework can find and segment objects without prior knowledge. Approaches described in [Campbell 07, Lai 12a, Karpathy 13, García 13] require a 3D scene model, which is a costly operation and can be unavailable to an autonomous robot due to multiple obstacles.

The concept of active segmentation or attention-driven segmentation was first presented by Aloimonos *et al.* [Aloimonos 88]. It was argued that the human visual system investigates and observes the scene by a set of fixations that are followed by segmentation. Attention-driven segmentation usually has two stages. During the first stage, a selection mechanism detects candidate object locations. During the second stage, the detected objects are segmented. The attention-driven segmentation approaches in [Ouerhani 01, Ko 06, Mishra 09, Björkman 10, Kootstra 10a, Bergström 11, Mishra 11] propose different solutions for the two stages. Given that the selection mechanism is directed by a specific search task (*i. e.* [Frintrop 06b]) in the case of robotic applications, attention-driven segmentation finds greater use than discriminative segmentation approaches. Interactive segmentations can be considered as a subset of attention-driven segmentations [Rother 04, Gulshan 10], with the attention mechanism performed by the user. Sedlacek and Zara [Sedlacek 09] presented an interactive point cloud segmentation approach based on minimizing Euclidean distance as energy function and employed user interactions to improve segmentation of the object of interest.

Ko *et al.* [Ko 06] proposed to segment objects of interest based on the principles of human attention and semantic region clustering. In [Ouerhani 01] Ouerhani *et al.* used a "seeded region growing" technique to segment an image w.r.t. seed points. Algorithms of [Ko 06, Ouerhani 01] are based on color images. However, since cheap and powerful active range sensors became available, a renewed interest in 3D methods was sparked throughout all areas of computer vision and robotics.

Mishra *et al.* [Mishra 09, Mishra 11] proposed an active segmentation algorithm based on provided seed points situated on objects. At first, in [Mishra 09] a probabilistic edge map, previously presented in [Martin 04], is calculated and improved using depth information or motion cues. This probabilistic edge map is remapped from the Cartesian coordinates to Polar coordinates with seed points (attention points) as poles, and the optimal cut through the polar edge map is found using the principle of minimization of path energy. This cut defines the border of an object. Though originally in the paper no strategy for detection of fixation points was proposed, it was discussed that visual attention mechanisms [Walther 06, Itti 98, Bruce 09] can be used for such a selection. The algorithm described in [Mishra 11] calculates attention points based on the principle of the boundary pixel border ownership. The idea of border ownership was described in detail by Zhou *et al.* [Zhou 00]. This means that each border knows to which object it belongs, and therefore a proper attention point inside the object can be selected. Kootstra *et al.* [Kootstra 10a] proposed an attention-driven graph-cut segmentation. Objects are localized with fixation points extracted from 2D symmetry saliency maps [Kootstra 08]. An energy minimization func-

tion is applied to depth and color along with a support plane constraint for segmentation. It is worth mentioning that the above segmentation algorithms [Kootstra 10a, Mishra 11] were developed specifically for table top scenes and require information about the support plane, which significantly limits their usage in real world applications. Bjorkman *et al.* [Björkman 10] presented an approach for attention-driven segmentation where fixation was used for unsupervised initialisation to generate object hypotheses using models of appearance, 3D shape, and size. The algorithm requires a-priori knowledge about the number of objects presented in the scene, which is not always possible in robotic scenarios. Bergstrom *et al.* [Bergström 11] proposed a segmentation algorithm based on the generation of segmentation hypotheses from 2D and 3D cues. The key idea is to obtain proper object segments through human-robot interaction. In our particular scenario, we want the robot to be completely autonomous, which rules out any human intervention.

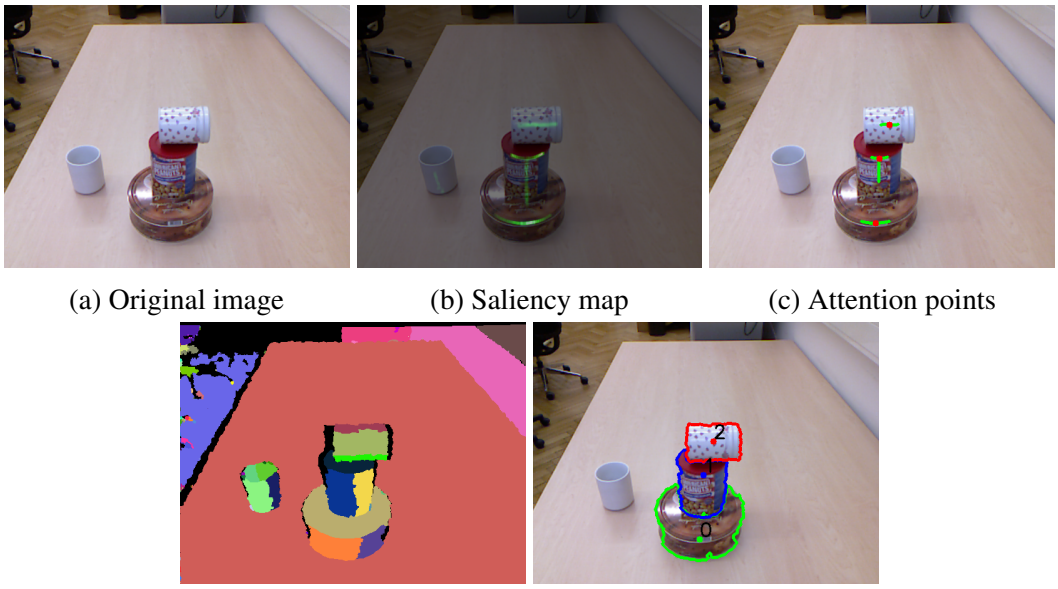
Fig. 3.7 and Fig. 3.9 show segmentation results using different state-of-the-art algorithms. As can be seen, current state-of-the-art attention-driven segmentations fail to segment objects if the scene consists of multiple occluded and cluttered objects, having several colors and textures. These types of scenes are common in domestic robotic tasks and need to be resolved correctly to enable manipulation of objects.

3.2 Attention-driven Segmentation for Domestic Scenes

In this Section, we present two novel attention-driven segmentation algorithm for cluttered table scenes. Our segmentation methods consist of two stages: object detection and segmentation. Object detection by extracting attention points from saliency maps was explained in Chapter 2. We introduce a novel selection of attention points algorithm from 3D symmetry-based saliency maps [Potapova 12b] to use them as fixations for the subsequent object segmentation. Furthermore, two novel segmentation algorithms are proposed. The first approach is called Compact Object Segmentation (COS) [Potapova 14b] and uses attention points to enable clustering of planar surface patches, similar to [Richtsfeld 12], using color similarity and the notion of compactness. The second approach is called Edge-based Segmentation (EBS) [Potapova 12a] and follows the idea proposed by Mishra *et al.* [Mishra 09], but instead of using a probabilistic edge map [Martin 04], we introduce a new type of edge map obtained by probabilistic learning.

We evaluated our approach on two publicly available datasets (Object Segmentation Database (OSD) [Richtsfeld 13] and Washington RGB-D Object Dataset (WOD) [Lai 12b]). The datasets contain different types of indoor scenes ranging from simple to complex scenarios. Our methods show improvements in terms of segmentation quality compared to existing attention-driven segmentation algorithms [Mishra 09, Mishra 11, Gulshan 10]. As can be seen from the picture (Fig. 3.7), the proposed approach successfully overcomes difficulties that other state-of-the-art algorithms experience when segmenting colorful and textured objects in highly cluttered scenes, commonly encountered in domestic environments .

The rest of the Section is organized as follows: In Section 3.2.1 we describe the proposed algorithms in detail. Section 3.2.4 shows the comparison of the proposed algorithms



(a) Original image (b) Saliency map (c) Attention points

(d) Surface patches (e) Segmentation result

Figure 3.1: *Object detection process:* starting from original image (a), 3D symmetry-based saliency map is calculated (b) (shown in green color on the original image); (c) shows attention points from symmetry (red) and skeletal line segments (green) (please note that for visualization purposes both attention points and skeletons were dilated); (d) shows planar surface patches and (e) shows segmentation result with respective attention points.

to other attention-driven segmentation algorithms.

3.2.1 Object Detection and Segmentation

In this Section, we describe the detection of good object candidate locations in a cluttered scene with further segmentation. Einhäuser *et al.* [Einhäuser 08] showed that objects attract human attention better than early vision saliency features. Symmetry is one of the characteristics of many natural as well as human-made objects and at the same time a powerful attentional cue [Kootstra 08, Potapova 12b]. Therefore, we based our object detection strategy on the calculation of a 3D symmetry-based saliency map. The algorithm to calculate 3D symmetry-based saliency map and extract attention points $\{f_k\}$ by means of T-junctions was described in Sections 2.3.4 and 2.5.3.

Particularly, for a given saliency map S we extract salient regions as 8-neighbor connected components of pixels with saliency value larger than zero $\{C_k\}$. The average saliency $\overline{S}_k$ of each salient region C_k is computed as

$$\overline{S}_k = \frac{1}{n_k} \sum_{p \in C_k} S(p) \tag{3.1}$$

where n_k is the number of pixels in the region, $S(p)$ is the saliency value of the point p.

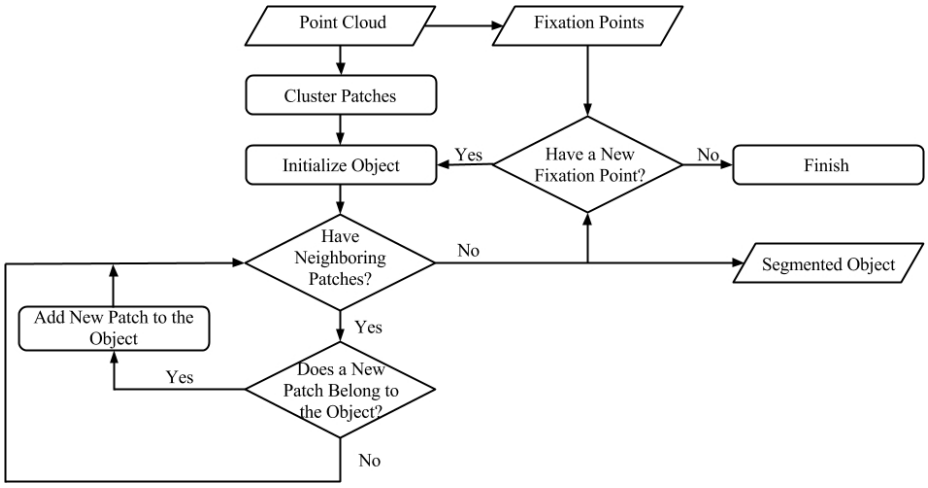


Figure 3.2: A general scheme for Compact Object Segmentation (COS).

Small errors in normal calculation due to sensor noise accordingly introduce noise in the saliency map. To this end we remove salient regions C_k where $\overline{S}_k < \theta_{sal}$. In our experiments we set θ_{sal} to 10% of the maximum saliency value. A better noise model for the sensor can be used to eliminate the need for this threshold.

The skeleton T_k is extracted from the connected component C_k. Symmetry attention points $\{f_k\}$ are extracted from the skeleton T_k as junction points, if they exist, or as mid-points for simple skeletal line segments. At the last stage attention points $\{f_k\}$ are sorted according to the average saliency $\overline{S}_k$ of the connected component C_k they have been extracted from. Figure 3.1c shows examples of attention points $\{f_k\}$ and skeletons T_k.

Given attention points $\{f_k\}$, we want to segment objects in the scene.

3.2.2 Compact Object Segmentation (COS)

The compact object segmentation algorithm was designed to segment compact objects that are typically present in household environments. The algorithm consists of several steps (Fig. 3.2). We first cluster points into planar patches based on their normals similar to Richtsfeld *et al.* [Richtsfeld 12]. We then cluster these patches beginning from the attention points by connecting similar patches as long as a given objectness measure is valid. The algorithm pipeline is shown in Fig. 3.1

Clustering Normals

Neighboring points are clustered to uniform patches without discontinuities using point normals. Normal clustering starts at the point with lowest curvature and greedily assigns neighboring points as long as they fit to the initial plane model. The algorithm iteratively creates planar surface patches until all points belong to some plane or are identified as noise. After normal clustering we obtain a set of planar surface patches $\{\rho_t\}$ (Figure 3.1d).

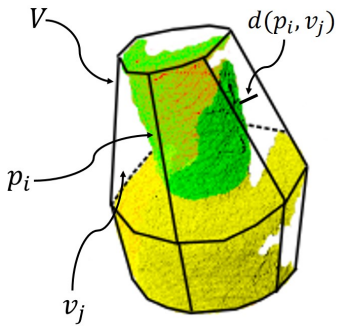


Figure 3.3: *Compactness measure:* green points represent the object candidate μ and yellow points represent a new patch ρ', that we want to add to the object candidate; p_i, V, v_j and $d(p_i, v_j)$ represent respectively a point from $\{\mu \cup \rho'\}$, convex hull, visible face in the convex hull and a distance from the point to the convex hull. In this particular case, compactness measure will result in high value, meaning that object is not compact.

Clustering Patches

Patches $\{\rho_t\}$ are now greedily clustered into object hypotheses $\{\mu_k\}$. Object hypotheses are initialized using symmetry attention points $\{f_k\}$. Attention points are sorted in decreasing order of saliency $\overline{S}_k$. Given a symmetry point f_k, all patches ρ_t bordering this point (with a 5 pixel radius) form an initial cluster. Patches are then greedily added to the cluster subject to a color and compactness constraint. Once a cluster cannot be extended further, the next cluster is initiated from the next attention point.

Color Similarity

A new patch ρ'_t is considered to be a part of the object only if its color model is similar to the already existing model for the object. The color similarity CS_{in} between a new patch ρ'_t and an object μ_k is computed as the correlation distance between their HSV color histograms.

$$CS_{in}(\mu_k, \rho'_t) = corr(H(\mu_k), H(\rho'_t)) \tag{3.2}$$

Object Compactness

A new patch patch ρ'_t is only added to an object, if its addition does not violate the compactness measure κ_k of the object (Fig. 3.3). Compactness κ_k is calculated as the mean of the shortest distances of the object points to the visible surfaces of the object's 3D convex hull. Let a set $\{p_{ki}\}$ be object points, V_k be the corresponding object's convex hull, and $\{v_j\}$ be a set of faces facing the viewpoint. Compactness measure κ_k is then calculated as:

$$\kappa_k = \frac{1}{n_k} \sum_{p_{ki}} d_{min}(p_{ki}, V_k) \tag{3.3}$$

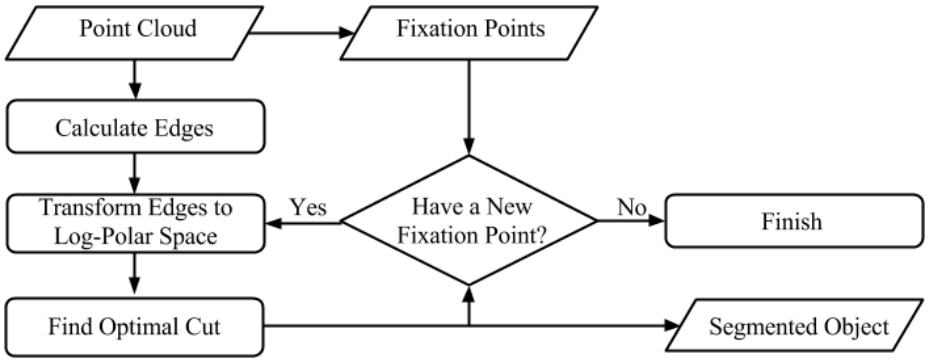


Figure 3.4: A general scheme for Edge-based Segmentation (EBS).

where n_k is the number of object points and $d_{min}(p_{ki}, V_k)$ is the shortest distance from the point to any visible face

$$d_{min}(p_{ki}, V_k) = \min_j d(p_{ki}, v_j) \tag{3.4}$$

Given two features, namely *color similarity* and *object compactness*, we train a support vector machine (SVM) [Chang 11] using *linear kernel* on the training part of the OSD to classify between new patches ρ'_t that belong to the given object hypothesis μ_k, and those that do not.

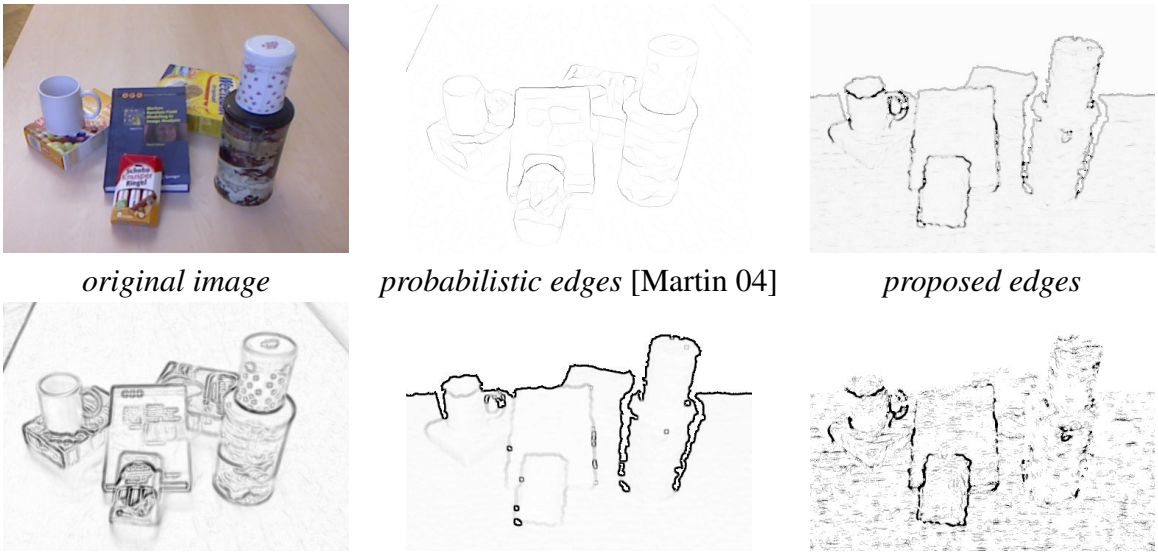
Examples of segmented objects using these constraints are shown in Fig. 3.1e and Fig. 3.7.

3.2.3 Edge-based Segmentation (EBS)

Probabilistic Edges for Segmentation

In 2009, Mishra *et al.* [Mishra 09] proposed to segment an object by transferring the edge map of the image from Cartesian coordinates to Polar coordinates with the attention point on the object acting as a pole and finding the best cut (Fig. 3.4). Segmentation results of the algorithm strongly depend on the quality of edges. The authors proposed to use edges of [Martin 04] weighted accordingly to disparity or motion cue information. In our scenarios, however, the segmentation approach did not show promising results (Fig. 3.7). This can be explained by the complexity of the object configurations in the scenes. Probability edges computed using [Martin 04] are based on color, intensity and texture. The algorithm is computationally expensive and 3D depth information is not explicitly used in edge calculation.

To improve segmentation results and speed up the computation we incorporated available 3D and 2D cues to obtain edges more suitable for our scenario. In our approach, we combine three different types of edges: **Sobel Color Edges (SC)** are calculated from color images; **Sobel Depth Edges (SD)** are computed from depth images using the Sobel operator, and indicate jump (occluding) edges; **Curvature Edges (CE)** are computed using information about point normals, as curvature discontinuity is often an indicator for object boundaries. Curvature discontinuities are calculated as the difference between a normal



original image | *probabilistic edges* [Martin 04] | *proposed edges*

sobel color edges (SC) | *sobel depth edges (SD)* | *curvature edges (CE)*

Figure 3.5: *Visual comparison of different types of edges*. Row 1 shows original image, probabilistic edges [Martin 04] and proposed edges respectively. Row 2 shows sobel color edges (SC), sobel depth edges (SD) and curvature edges (CE). As can be seen, the proposed algorithm to compute edges gives more visually pleasing results than probabilistic edges [Martin 04].

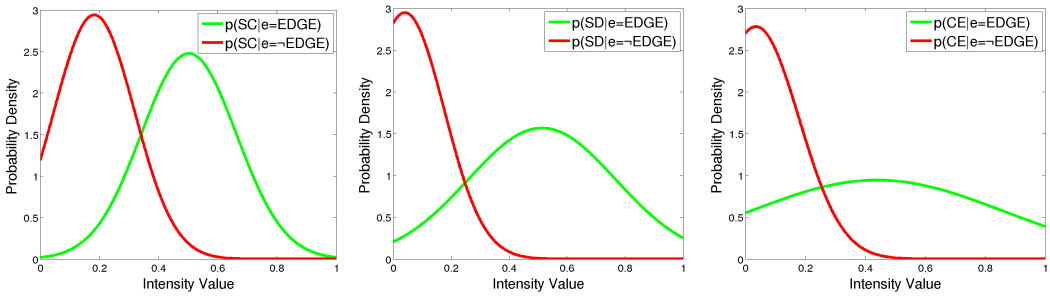
of a point and the averaged normal over its neighborhood. Examples of different types of edges are shown in Fig. 3.5.

To fuse the above types of edges into single edge map, we first learn the two probability distributions $p\left(c|E=edge\right)$ and $p\left(c|E=\overline{edge}\right)$ for each edge type $c \in \{\text{SC}, \text{SD}, \text{CE}\}$ (Fig. 3.6). The better the separation between the positive and negative distributions, the better a given observation distinguishes between edges and non-edges. Assuming that all three edge observations (SC, SD, CE) are conditionally independent, the probability of E being an edge is then given by the Bayes rule as:

$$\begin{aligned} & p\left(E=edge|SC,SD,CE\right) = \\ & = \frac{p\left(E\right)\prod_{c\in\{SC,SD,CE\}} p\left(c|E\right)}{\prod_{c\in\{SC,SD,CE\}} p\left(c|E\right) + \prod_{c\in\{SC,SD,CE\}} p\left(c|\overline{E}\right)} \end{aligned} \tag{3.5}$$

We assume that the a-priori probability of E being or not being an edge is equally likely $p\left(E\right) = p\left(\overline{E}\right) = 0.5$. Learning was done on the training part of the OSD [Richtsfeld 13]. An example of a final probabilistic edge map is shown in Fig. 3.5.

To segment an object using an attention point f_k, the probabilistic map is transformed from the Cartesian coordinate space to Polar coordinate space with f_k as a pole, as in [Mishra 09]. Object borders are found as the optimal cut of the edge map in the Polar coordinate space. Segmentation examples are shown in Fig. 3.7.



(a) Sobel Color Edges (SC) (b) Sobel Depth Edges (SD) (c) Curvature Edges (CE)

Figure 3.6: *Probability density functions (PDFs)* for different types of edges.

3.2.4 Results and Evaluation

We evaluated our segmentation algorithm on two publicly available databases: Object Segmentation Database (OSD) [Richtsfeld 13] and Washington RGB-D Object Dataset (WOD) [Lai 12b] Other databases as Caltech256, Pascal VOC, LabelMe, Berkeley's B3DO, NYU's Depth Dataset, UW's RGB-D Object Dataset do not cater to our specific task of cluttered table scene segmentation.

OSD database is targeted towards segmentation evaluation and consists of scenes with varied object configuration complexities. It is composed of images with complex and cluttered scenes, as well as scenes where only several boxes or other simple objects are presented, as shown in Fig. 3.7. The OSD database consists of training and testing parts, and objects in the dataset were hand-labeled with polygon outliers.

WOD database consists of several sequences of scenes recorded in typical indoor scenarios. Ground truth in the dataset is offered by means of bounding boxes indicating object presence. Since we are interested in segmentation in terms of object boundaries, bounding boxes are not suitable for evaluation. Therefore we randomly selected 87 images from WOD evenly distributed among all sequences. Image examples are shown in Fig. 3.7. Objects in these 87 images were hand-labeled by approximating object boundaries with polygons.

Evaluation was carried out by considering two specific aspects – namely, object detection (*i. e.* placing attention points on objects) and object segmentation. We do not attempt to directly measure the quality of object detection, but instead present the effect of choices in the object detection methodology on our object segmentation approach. In addition, the evaluation was performed to compare the performance of our approach against several state-of-the-art segmentation approaches.

Object Detection Strategies

In our work, we applied several strategies for object detection in order to estimate their influence on the object segmentation. As described earlier, the primary object detection strategy used in our pipeline involves the generation of saliency maps from 3D symmetries. In this strategy (TJ3D), attention points are selected as points of T-Junctions (or midpoints for simple lines) in symmetry lines extracted from the 3D symmetry-based saliency

Figure 3.7: *Visual comparison of different segmentation algorithms*. Column 1-2 show segmentation results for OSD [Richtsfeld 13] dataset and columns 3 shows segmentation results WOD [Lai 12b] dataset. Segmentation masks and attention points are shown in different colors with respective numbering reflecting the order of attention shift. Results for our COS and EBS algorithms are shown in the last two rows. As can be seen, all algorithms except the proposed approach have difficulties handling cluttered table scenes.

Segmentation	OSD [Richtsfeld 13]					
	All		Best		First	
	Mean	Std	Mean	Std	Mean	Std
G10	0.49	0.07	0.52	0.07	0.54	0.06
M09+TJ3D	0.56	0.07	0.57	0.07	0.59	0.06
M09+WTA2D	0.50	0.07	0.50	0.07	0.64	0.05
M09+WTA3D	0.53	0.07	0.54	0.07	0.64	0.05
M11	0.64	0.07	0.70	0.06	0.74	0.05
EBS	0.72	0.05	0.74	0.05	0.75	0.05
COS	**0.89**	0.02	**0.90**	0.02	**0.91**	0.01
Segmentation	WOD [Lai 12b]					
	All		Best		First	
	Mean	Std	Mean	Std	Mean	Std
G10	0.34	0.12	0.47	0.11	0.52	0.10
M09+TJ3D	0.61	0.09	0.64	0.09	0.69	0.07
M09+WTA2D	0.59	0.10	0.63	0.09	0.72	0.05
M09+WTA3D	0.60	0.10	0.63	0.09	0.71	0.06
M11	0.56	0.13	**0.74**	0.08	**0.79**	0.05
EBS	0.56	0.09	0.67	0.07	0.72	0.05
COS	**0.63**	0.08	0.72	0.06	0.74	0.05

Table 3.1: F-score for different segmentation algorithms evaluated on OSD [Richtsfeld 13] and WOD [Lai 12b] datasets.

maps (Fig. 3.1c). The second strategy (WTA3D) employed, extracts attention points using Winner-Take-All (WTA) [Lee 99] from the 3D symmetry-based saliency maps. To see how the use of 3D information improves the quality of a detection strategy, we also include attention points using Winner-Take-All from 2D symmetry-based saliency maps [Kootstra 08] (WTA2D).

Object Segmentation

To evaluate the attention-based aspect of the algorithm we performed comparison against an attention-driven active segmentation algorithm proposed by Mishra *et al.* [Mishra 09] (M09), as well as its extension which uses depth information as described in [Mishra 11] (M11).

Though it is clearly not fair to compare our approach to algorithms that use only color information, it is still interesting to see how the performance differs. Interactive segmentation algorithms [Rother 04, Gulshan 10, Boykov 01] require user input (*e. g.* bounding box as in [Rother 04]). In scenarios where it is not possible for a user to provide input, the user behavior can be simulated by a computational model of the visual attention system [Bruce 09, Borji 13]. Therefore, we selected a state-of-the-art interactive segmentation al-

gorithm presented by Gulshan *et al.* [Gulshan 10] (G10) to compare with our algorithm. The algorithm proposed in [Gulshan 10] requires strokes of foreground and background as input. Foreground strokes in our evaluation were simulated as twice dilated skeleton lines from saliency maps. Background strokes were simulated as rectangles near the image border 10% smaller than the size of the original image.

The output segmentation masks are compared to the ground truth labeling in terms of the F-measure defined as $2PR/(P+R)$. We calculated precision P as the fraction of the segmentation mask overlapping with the ground truth and recall R as the fraction of the ground truth overlapping with segmentation mask.

Segmentation algorithm M09 was evaluated using object detection strategies TJ, WTA2D and WTA3D, mentioned earlier. Segmentation algorithm M11 was evaluated with its own object detection strategy, because this strategy is an intrinsic part of the algorithm. Segmentation algorithm G10 was evaluated using symmetry lines as foreground strokes. Evaluation results are presented in Table 3.1.

Note that the attention mechanism cannot rule out that several attention points come to lie on the same object. In this case, each attention point leads to a possibly different segmentation for an object. Therefore, we calculated three F-scores: the label *first* in Table 3.1 refers to the segmentation from the first attention point, *best* refers to the best segmentation w.r.t. ground truth, and *all* refers to the average score over all segmentations for an object. If *first* is lower than *best* this means that the attention points are not optimal. Ideally the first attention point leads to the best segmentation. If *all* is significantly lower than *best* this means that the segmentation algorithm depends a lot on the position of the attention point. All F-scores in Table 3.1 are averaged over all objects and all scenes.

As can be seen from Table 3.1, the primary object detection strategy (TJ3D) proposed in this thesis results in improved performance for all types of segmentation algorithms compared to other detection strategies. Evaluation results also show that our combined approach of detection and segmentation performs better on both databases than state-of-art segmentation algorithms. Results for G10 show that color-only segmentation cannot handle complicated table scenarios without good user input. Fig. 3.7 shows visual segmentation outputs for different segmentation strategies. Segmentation results show examples when different attention points can give different segmentation results for the same object. It can be seen that the good results achieved by the proposed approach also correspond to visually pleasing segmentations compared to other attention-driven approaches.

3.3 Incremental Attention-driven Scene Segmentation

In this Section, we present an attention-driven incremental segmentation for RGB-D views of highly cluttered scenes inspired by the approach described in [Richtsfeld 12]. In contrast to [Richtsfeld 12], where the complete scene is segmented, our approach segments at first most salient objects. This strategy results in significant reduction of the computational complexity of the algorithm. We show that the proposed approach can deal with cluttered scenes in a more efficient manner compared to other attention-driven segmentation approaches, as well as output better segmentation results. The proposed attention-driven

Data: Point cloud $\mathbf{P} = \{\mathbf{r_1}, ..., \mathbf{r_N}\}$, saliency map SM
Result: Object segmentation $\overline{g}$
$S \leftarrow SegmentLowLevelsurfaces(\mathbf{P})$;
$S \leftarrow SortSurfaces(S, SM)$;
$S' \leftarrow \{s_1\}'$;
$\overline{g} \leftarrow \emptyset$;
while *true* **do**
  $S' \leftarrow ComputeSurfaceModels(S')$;
  $\Re \leftarrow ComputeSurfaceRelations(S')$;
  $W \leftarrow GetRelationWeights(\Re)$;
  $\overline{g}' \leftarrow GraphBasedSegmentation(S', W)$;
  if $\overline{g} \neq \overline{g}'$ **then**
    $S' \leftarrow S' \bigcup Neighbors(S')$;
    $\overline{g} \leftarrow \overline{g}'$
  else
    $return(\overline{g})$;
  end
end

Algorithm 1: Attention-driven Segmentation

mechanism enables quick response and linear degradation in time performance with the increase of the scene complexity. Scene complexity is defined as the degree of the object clutter (Fig. 3.13). Incremental segmentation means the most relevant (salient for the given task, *i. e.* having required color) objects are segmented first, and the segmentation proceeds in order of decreasing visual saliency. Starting with the most salient part, the method successively modifies the current object hypothesis using perceptual grouping principles. An object is considered to be segmented when no more additional information about the object can be extracted from the scene. Up to our knowledge this is the first algorithm that combines benefits of both attention-driven and incremental scene segmentation.

We evaluated our approach on two publicly available datasets (Object Segmentation Database (OSD) [Richtsfeld 13] and RGB-D Object Dataset (WOD) [Lai 12b]). These datasets contain different types of indoor scenes ranging from simple to complex scenarios. Our method shows improvements in terms of segmentation quality and computational performance compared to the existing attention-driven segmentation algorithms [Mishra 09, Kootstra 10a, Mishra 11].

3.3.1 Overview of the Object Segmentation

In this Section, we give an overview of the proposed algorithm, with a subsequent detailed description of each step of the algorithm as depicted in Alg. 1. The algorithm starts by segmenting a point cloud into surfaces based on point normals similarity. Given a saliency map of the scene, the saliency of each surface is computed. The object of interest segmentation is initiated by selecting the most salient surface. The following steps are then performed iteratively: selected surfaces are fitted with parametric models and grouped us-

ing graph-based segmentation; neighboring surfaces are selected and the process repeats. Segmentation stops when no more additional information about the object can be extracted, *i. e.* adding new surfaces does not change the segmented object candidate obtained during the previous iteration.

Low-Level Segmentation

The low-level segmentation is based on adaptive clustering of points in the point cloud $\mathbf{P}$, where a point $\mathbf{r_i} \in \mathbf{P}$, $\mathbf{r}_i = (r_i, c_i, \mathbf{p}_i, \mathbf{n}_i)$, (r_i, c_i) are image pixel coordinates, $\mathbf{p}_i = (x, y, z)$ are 3D space coordinates and $\mathbf{n}_i = (n_{ix}, n_{iy}, n_{iz})$ is the estimated normal vector at this point. A point is added to a surface if (1) its normal deviation from the average surface normal and (2) its distance to the surface are smaller than given thresholds (these thresholds are adapted), based on the depth value of the respective point to take into account noise behavior of the depth sensor [Mörwald 13]. Low-level segmentation outputs a set of surfaces $S = \{s_1, ...s_n\}$.

Surface Sorting

Given a saliency map SM we calculate the saliency SM_j of a surface s_j:

$$SM_j = \frac{\sum_{\mathbf{r_i} \in s_j} SM(r_i, c_i)}{|s_j|} \tag{3.6}$$

where $|s_j|$ is the size of the surface s_j, *i. e.* the number of points. Surfaces are sorted according to their saliency values. The most salient surface initializes the set of selected surfaces S'.

Surface Modeling

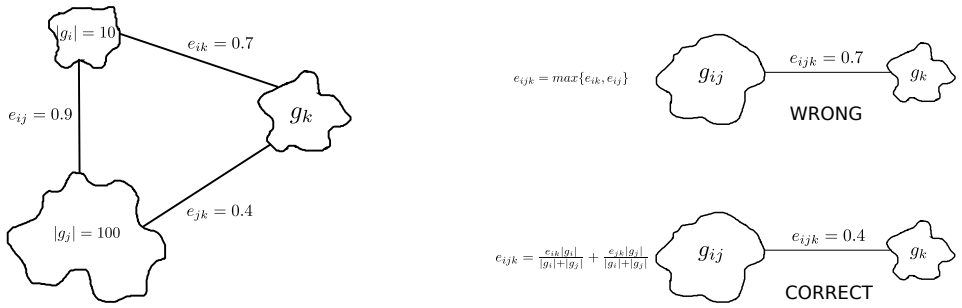
Given a set of selected surfaces $S' = \{s'_{k_1}...s'_{k_n}\}$, $S' \in S$, we model each surface s'_{k_i} using either a plane or a NURBS [Mörwald 13]. A surface s'_{k_i} is modeled with a NURBS if the data is better explained using this higher-parametric model in terms of a minimum description length (MDL) criterion, taking into account number of explained points, model complexity, and fitting error. Neighboring surfaces (s'_{k_i}, s'_{k_j}) are fused and modeled using a single NURBS $s''_{k_i} = s'_{k_i} \cup s'_{k_j}$ if again the combined model is better in terms of MDL than the two individual models.

Surface Relations

After modeling individual surfaces in the set S', we define a set of relations between pairs of neighboring surfaces. These relations model perceptual grouping principles, such as similarity or good continuation, and serve to identify groups of surfaces that together make up a single object. This set of relations $\Re$ defines a feature vector $\mathbf{r_{ij}}$ for each pair of neighboring surfaces (s'_i, s'_j). The feature vector includes such relations as color similarity; Gabor and Fourier texture similarities; relative size; color, curvature and depth similarities and variations on the border; relation between 2D/3D borders. All features are normalized

to be in the range $\langle 0, 1 \rangle$. A support vector machine (SVM) [Chang 11] was trained to classify between same/different object given a feature vector. Training was done on the ground truth annotated scenes in the learning part of the Object Segmentation Database (OSD) [Richtsfeld 13] using 10 fold cross-validation. The best classification accuracy is achieved using an *RBF* kernel with $C = 512$ and $\gamma = 2$. For every feature vector $\mathbf{r_{ij}}$ the SVM then outputs the confidence score w_{ij}, which is further used as the weight between surfaces s'_i and s'_j.

Graph-Based Segmentation



a) graph *before* merging groups g_i and g_j *b)* graph *after* merging groups g_i and g_j into group g_{ij}

Figure 3.8: *Graph-based segmentation*. a) shows a graph with three groups g_i, g_j and g_k before merging groups g_i and g_j. As shown in the drawing, groups g_j and g_k are weakly connected, while groups g_i and g_k connected strongly. b) shows resulting graphs after merging for two different strategies: the upper drawing shows the strategy that does not take into consideration the number of actual points in the groups, and the lower drawing shows the proposed strategy which takes into consideration the actual number of points in the groups. As can be seen from the drawing the proposed strategy gives the correct result. The other one will eventually lead to groups g_{ij} and g_k being merged together.

Given a set of selected surfaces S' and their pairwise weights W, we form an undirected graph $G = \langle V, E \rangle$, where vertices V are selected surfaces S' and edges E with weights W. To segment the graph we use the method proposed by Felzenszwalb *et al.* [Felzenszwalb 04]. The algorithm merges vertices of the graph into groups $G = \{g_i\}$. In our case each surface is counted as one element in the group.This means that when two groups g_i and g_j (with edge weight e_{ij}) are going to be merged together and there exists a third group g_k weakly connected to g_i (with weight e_{ik}) and strongly connected to g_j (with weight e_{jk}) then eventually g_k will be merged with g_{ij} with no respect to the the actual number of points in each group (Fig. 3.8). To prevent this we modify edge weights e_{ijk} between g_k and new merged group g_{ij} as follows:

$$e_{ijk} = \frac{e_{ik}|g_i|}{|g_i| + |g_j|} + \frac{e_{jk}|g_j|}{|g_i| + |g_j|} \tag{3.7}$$

where $|g_i|$ and $|g_j|$ are cardinalities of respective groups, *i. e* . number of actual points.

Group $\overline{g}'$ containing the most salient surface represents the segmentation of the object of interest.

Object Creation

After graph-based segmentation, we check if the current segmentation $\overline{g}'$ of the object of interest has changed with respect to the segmentation $\overline{g}$ at the previous iteration. If the segmentation $\overline{g}'$ has not been changed, it is returned as a segmented object of interest. Otherwise, all neighbors of S' are added to S' and the process continues. Surfaces forming the segmented object of interest are not involved in segmentation of further objects. Please note that the complete segmentation is a special case of the incremental attention-driven segmentation, when all surfaces have been selected.

3.3.2 Evaluation

We have evaluated our algorithm on the two datasets: the Object Segmentation Database (OSD) [Richtsfeld 13] and the Washington RGB-D Object Dataset (WOD) [Lai 12b]. Objects in the database were hand-labeled by approximating object outlines with polygons. WOD consists of image sequences captured from 8 different scenes. We randomly selected 86 different images from all 8 scenes. All objects in each scene were also hand-labeled with polygons. The SVM classifier [Chang 11] for predicting the weight between two surfaces given their relations was trained on the training part of the OSD. Please note, that the SVM classifier was not re-trained to be used with WOD. This together with good segmentation quality show that the proposed approach generalizes well beyond its training set.

3.3.3 Segmentation Quality

To evaluate the robustness of the proposed approach against the location of attention points on the object we generated a set of sequences of random attention points and simulated saliency maps by placing a Gaussian at each attention point. For each object in each image we created five sequences of attention points. Each sequence consists of five attention points. The first attention point is always on the respective object in the image and the other four are on randomly selected objects (objects were sampled without replacement).

We compared the proposed algorithm (INC) to three other attention-driven segmentation algorithms: Mishra *et al.* [Mishra 09] (M09), Kootstra *et al.* [Kootstra 10a] (K10) and Mishra *et al.* [Mishra 11] (M11). In the segmentation algorithm proposed by Mishra *et al.* [Mishra 11], the choice of attention points is a significant part of the algorithm and cannot be replaced. This limits the usage of this segmentation in scenarios where a particular object should be segmented. Note also that K10, M09 and M11 require knowledge of the supporting plane, while the proposed approach does not. This significantly restricts the applicability of algorithms K10, M09 and M11 in real world tasks.

To evaluate the quality of segmentation, we used precision P and recall R metrics. Precision was calculated as the fraction of the segmentation mask overlapping with the ground truth and recall as the fraction of the ground truth overlapping with the segmentation mask. We averaged P and R for each object over all appearances of the object in sequences. We

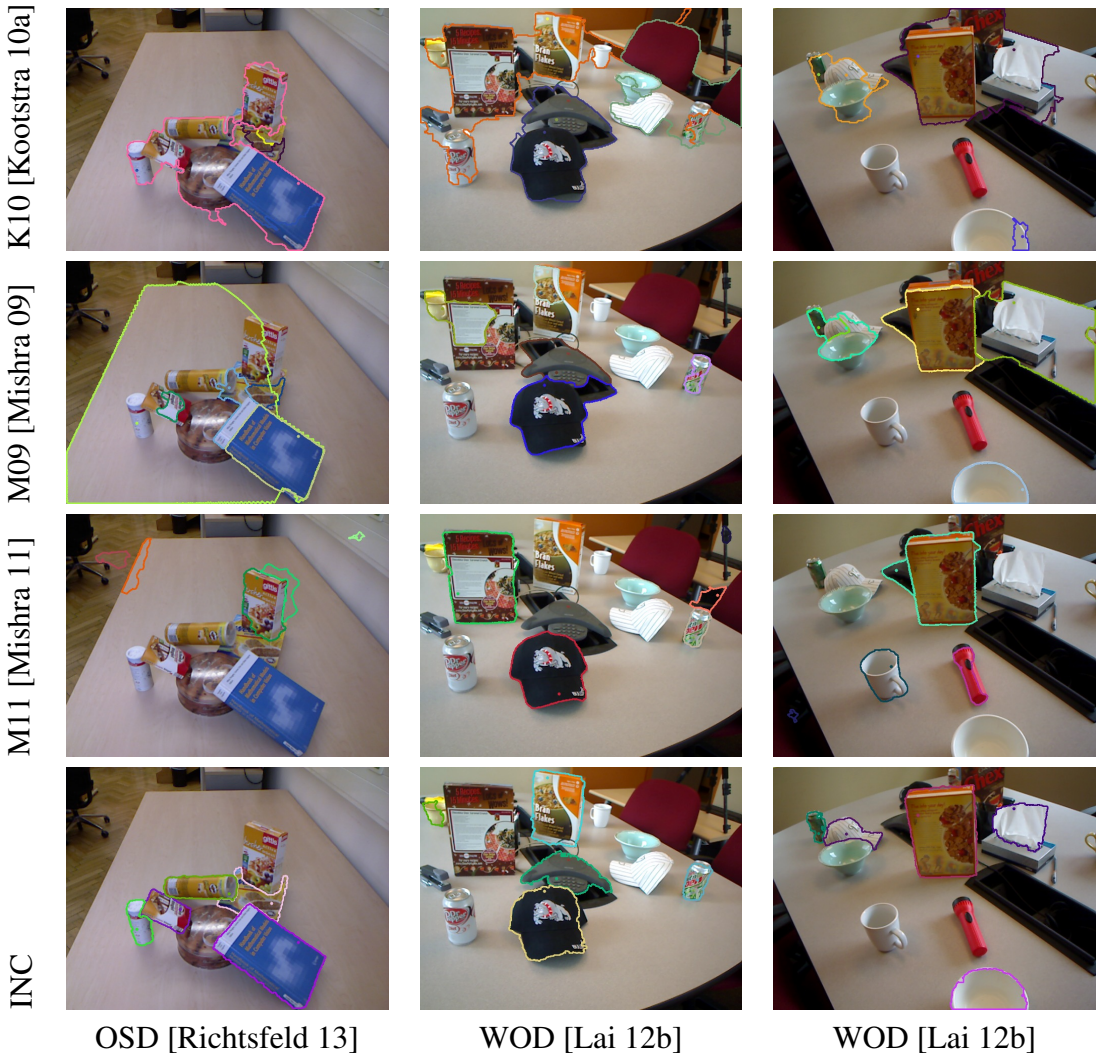
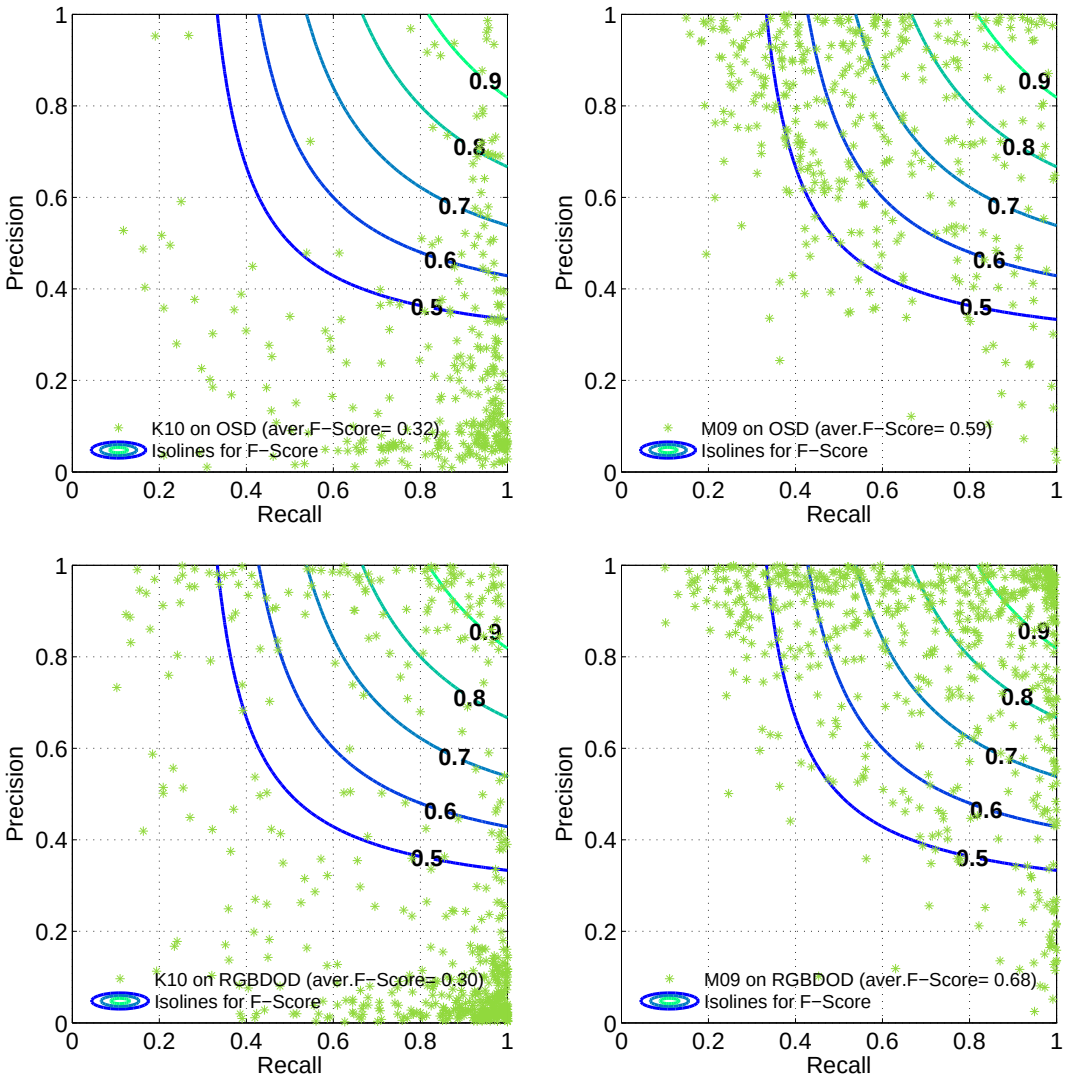


Figure 3.9: *Visual comparison of different segmentation algorithms.* Column 1 and columns 2-3 show segmentation results for OSD [Richtsfeld 13] and for WOD [Lai 12b] respectively for different segmentation algorithms. Corresponding segmentation masks and attention points are shown in the same color. As can be seen, existing algorithms have difficulties handling cluttered table scenes, while the proposed algorithm clearly performs well.

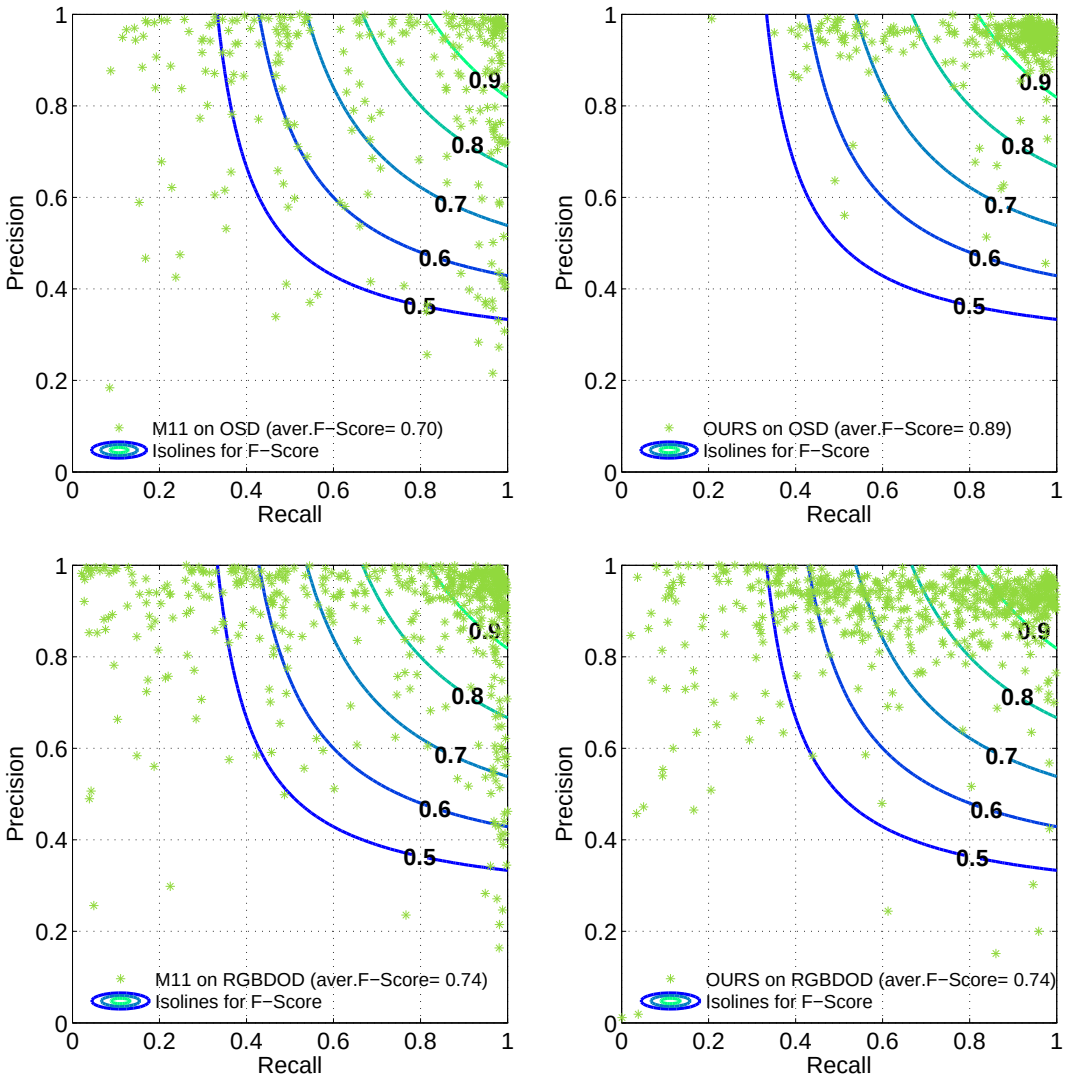
also provide the F-measure defined as $2PR/(P+R)$, averaged over all objects. The results are shown in Fig. 3.10 and Fig. 3.11. We compare the performance of attention-based segmentation to the full scene segmentation. Full scene segmentation is a particular case of attention-based segmentation, when all surfaces are selected. Please note, that the results of these two segmentations are not necessarily equivalent, because of the merging procedure. As can be seen, the proposed algorithm perfectly handles the complexity of the OSD, while the other attention-driven segmentation algorithms show lower performance in terms of F-score. The performance of the proposed algorithm visually looks slightly lower for



K10 [Kootstra 10a] M09 [Mishra 09]

Figure 3.10: *Precision vs. Recall Cues*. Column 1 and Column 2 show Precision vs. Recall of the segmentation for K10 [Kootstra 10a] and M09 [Mishra 09] segmentation algorithms for OSD [Richtsfeld 13] and WOD [Lai 12b] respectively. Each point in the plot represents averaged values of *Precision* and *Recall* for each object instance appearing in images.

WOD than for OSD. The reason is missing depth data in some images in the WOD. The proposed algorithm relies on the depth information and it is not possible to segment parts of the object, where there is no data (like black shiny objects or transparent bottles). Therefore, segmentation recall for these objects can be low. Performance of the K10 algorithm seems to be inferior, which can be caused by the choice of parameters. We used standard choice of parameters as set by the authors in their implementation.

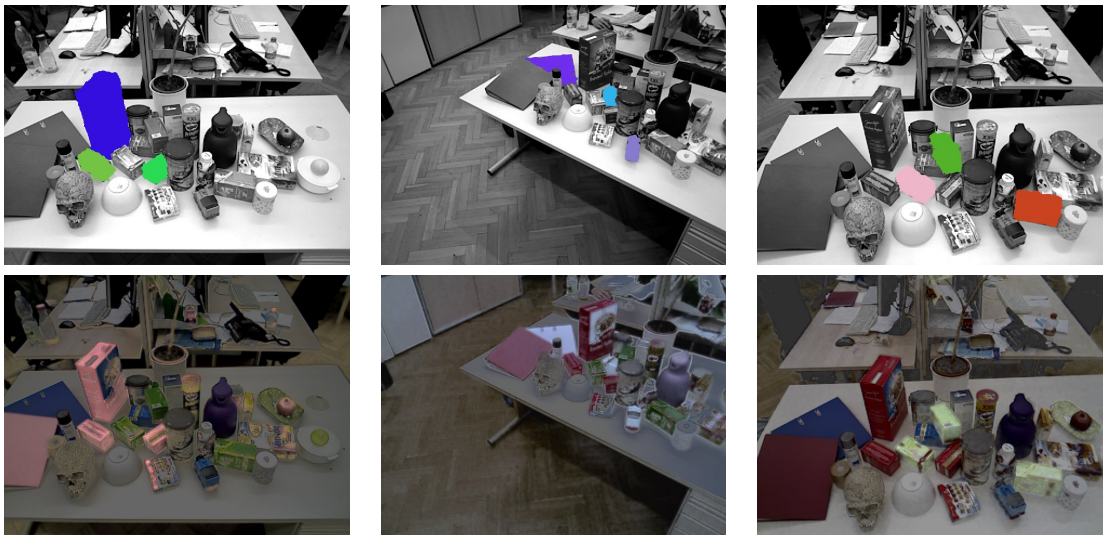


M11 [Mishra 11] INC

Figure 3.11: *Precision vs. Recall Cues*. Column 1 and Column 2 show Precision vs. Recall of the segmentation for M11 [Mishra 11] and proposed segmentation algorithms for OSD [Richtsfeld 13] and WOD [Lai 12b] respectively. Each point in the plot represents averaged values of *Precision* and *Recall* for each object instance appearing in images. As can be seen in combination with Fig. 3.10 the proposed algorithm outperforms other attention-driven segmentation algorithms in terms of $F - score$.

Qualitative Results

To demonstrate that incremental attention-driven segmentation can be successfully used in indoor environments with different types of saliency maps, we have run the proposed segmentation on video filmed in an indoor environment. Saliency maps based on color were used to direct segmentation.



Red Objects *Blue Objects* *Green Objects*

Figure 3.12: *Results of the proposed incremental attention-driven segmentation algorithm for the indoor video sequence*. Row 1 shows resulting segments in different colors. Row 2 shows saliency maps overlaid with original images where desired object are highlighted. Columns 1-3 correspond to segmentation of first three red, blue and green objects.

3.3.4 Computational Complexity

We further evaluated the runtime behavior of our method in comparison to M09 and M11. Note that as M11 is only available as a binary implementation, we could only measure the accumulative time of the segmentation of all selected objects. We do not compare to the time performance of K10 due to the inferior quality of the segmentation.

Time performance results are shown in Fig. 3.13. Images are ordered along the horizontal axis in increasing scene complexity. Since providing an objective measure of complexity is difficult, we defined complexity for the purpose of this evaluation as the time it took the original approach in [Richtsfeld 12] to segment the complete scene (hence the "complete" curve in the upper row of Fig. 3.13 is rising monotonically). This allows us to clearly show the runtime behavior w.r.t. complexity. Note however that this notion of complexity is tied to [Richtsfeld 12] and this ordering of images does not correspond to the complexity of methods M09 and M11, although a similar trend can be observed. Runtimes for each image are averaged over all runs over the images (i.e. all attention point sequences for all objects in the image). The top row of Fig. 3.13 shows that runtime for incremental attention-driven segmentation grows significantly slower with increasing complexity than when segmenting the whole scene and remains almost constant for segmenting the first object. Peaks and valleys in the plots stem from the complexity of the objects themselves and NURBS fitting consumes the majority of runtime. Thus, images containing objects with mainly flat surfaces, such that planes rather than NURBS are fitted, need significantly less time and represent the valleys in the curves. The bottom row shows runtimes of M09

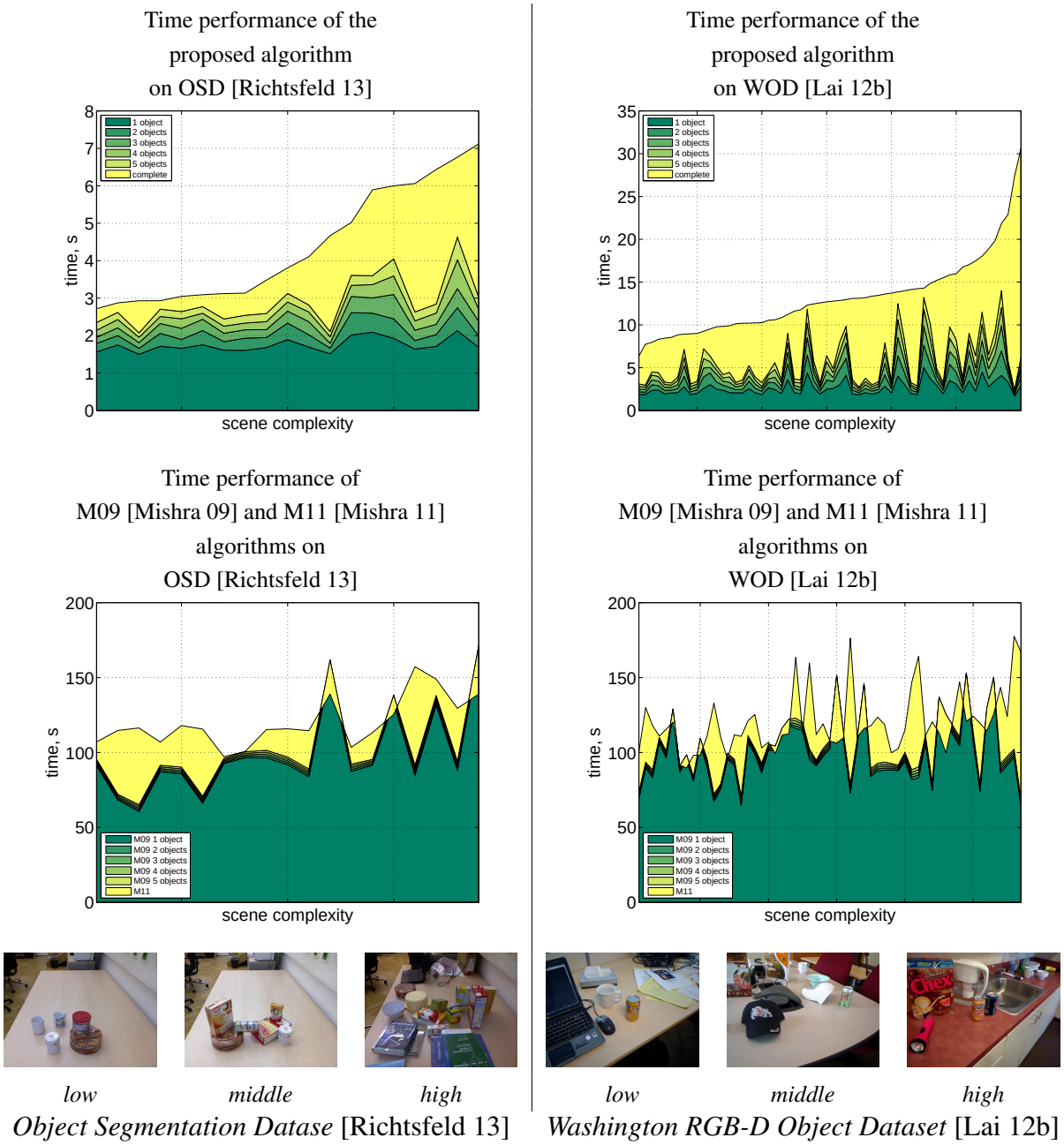


Figure 3.13: *Computational Complexity vs. Scene Complexity*. Row 1 and row 2 show time performance of different segmentation algorithms on OSD [Richtsfeld 13] and WOD [Lai 12b] respectively. Note that images are ordered along the horizontal axis in increasing scene complexity.

and M11. Both of these rely on a time-consuming 2D edge detection step, which explains the high offset of over 50 s. While M09 and M11 are generally slower than the proposed approach, they show the benefits of attention-driven segmentation too. Our algorithm is

implemented in *C++*, as are M09 and M11. Tests were run on an Intel Core i7 CPU at 2.93GHz.

3.4 Discussion

In this chapter we presented two novel attention-driven algorithms for cluttered table scene segmentation: Compact Object Segmentation (COS) and Edge-based Segmentation (EBS). Both methods incorporate a novel object detection strategy using a saliency operator based on 3D symmetry with attention point estimation based on symmetry lines and T-Junction points. The first method (COS) is based on greedy clustering of planar surface patches using the notion of compactness and color similarity The second method (EBS) utilizes a probabilistic framework to calculate probabilistic edges, based on color, depth and curvature. Our approach shows good results on typical cluttered table scenes containing human made objects. We have shown that our selection of attention points improves performance of attention based segmentation methods and that our combined attention and segmentation approach improves over state-of-the-art attention-driven segmentation approaches.

We also presented an incremental segmentation approach and showed on two challenging datasets how the computational complexity of finding the first few relevant objects in the scene is significantly reduced compared to complete scene segmentation. We found that the first object can be found in almost constant time, irrespective of the scene complexity. Furthermore, the proposed incremental approach outperformed existing attention-based segmentation in terms of segmentation quality and time performance. These results highlight the importance of providing the proper attention mechanisms, i.e. derived from the task context the robot performs in and the role vision plays in it.

In summary, we argue that the system view is more important than tuning specific processing routines in isolation. We argue that this is in general the only way forward: meeting (soft) real time constraints using anytime algorithms that are able to output some result at any time and more/better results with more time. Our approaches w.r.t. segmentation in this direction are attention-driven and incremental segmentations, starting at the most task relevant parts of the scene as highlighted by suitable attention mechanisms.

Chapter 4

Conclusion

To recognize or manipulate objects, we need to have their precise localization. In this thesis, to acquire precise location of objects we developed an attention-driven system for object detection and segmentation. Objects are detected using novel 3D saliency operators by means of extracting attention points. Then a state-of-the-art technique to segment detected objects using attention-driven algorithms is presented.

4.1 Summary

Object detection is a very important task and it plays a crucial role in robotics. In robotics applications, objects often appear in cluttered environments; moreover, they can often be occluded. Man-made objects are often multi-colored and multi-textured. All of this makes the problem of object detection highly non-trivial. In this thesis we proposed to detect object by means of extracting attention points from 3D saliency maps.

In robotics, 3D information is used in many areas such as navigation and grasping. While traditional attention models are mostly color-based, in this thesis we investigated the importance of 3D visual information for attention. To do that, we conducted a comprehensive analysis of eye-tracking experiments in 3D. We found that depth perception has a valuable role in human visual attention and hence robots will also benefit from incorporating 3D-based visual attention. To understand what is still missing in the field of 3D-based visual attention for robotics, we also studied existing attentional models. We found two main problems: (1) there are no models that involve 3D attention and that aim at solving such specific tasks as detecting suitable regions for grasping; (2) while research on eye-tracking suggests that objects are more attractive than single features, such methods are underrepresented in robotic vision.

To handle the first problem, we proposed a set of saliency cues, based on such properties as height and surface orientation. We showed that their fusion results in better object detection abilities.

To handle the second problem, we developed a novel 3D Symmetry-based saliency operator that works on partial object views. This proposed attentional operator was specifically designed to detect generic objects by detecting object symmetries in cluttered scenes.

Moreover, a new extraction strategy designed specifically for this 3D Symmetry-based saliency operator was proposed. Evaluation results showed significant increase in detection quality for generic objects, compared to existing state-of-the-art saliency maps and extraction strategies.

We exhaustively evaluated 159 different combinations of multi-level strategies, combination methods and normalization operators to create saliency maps for images in a publicly available database with cluttered table-top scenes. The goal of the evaluation was to understand if more sophisticated methods result in better performance in terms of object detection. We came to the conclusion that more complicated, and therefore more computationally expensive methods, do not improve detection performance under most scenarios.

To be able to use detected objects in tasks such as manipulation, we proposed to segment them using two different novel attention-driven algorithms. The first algorithm segments objects using an edge map based on color, depth and curvature within a probabilistic framework. The second segmentation algorithm is based on clustering locally planar surface patches using the notion of compactness and color similarity.

Moreover, we extended attention-driven segmentation to be able to segment as many objects as necessary, instead of only one, as in classical attention-driven segmentation algorithms. We called this extension incremental segmentation. Incremental segmentation detects and prioritizes the processing of objects of interest using attention, i.e. it segments the object of the highest importance first and then incrementally explores the rest of the scene based on the saliency of each region. Eventually, given enough time, incremental segmentation segments the complete scene. The evaluation on cluttered indoor scenes, containing man-made objects clustered in piles and dumped in a box, shows that our proposed methods outperform state-of-the-art methods of attention-driven segmentation in terms of segmentation quality and computational performance.

4.2 Outlook

In this thesis, we have seen how the proposed attention-driven framework increases the performance and the quality of object detection and segmentation. These encouraging results open the way to solving remaining challenges in object understanding, localization and representation that should be addressed in the near future. One of them is finding a more elegant way to combine both 2D and 3D attention models to gain better results on a variety of different types of scenes. The other one is in creating top-down attention operators that will lead to better detection results in specific scenarios such as finding unknown objects with known characteristics. Another challenge is to create attentional models that can direct attention to any object, regardless of its size or type. Attention-driven segmentation can be further improved by incorporating various shape cues, as well as developing a better edge model. Segmentation of more task specific scenes will benefit by adding more Gestalt-based features to increase the scope of categories of objects properly segmented.

Bibliography

[Achanta 08] R. Achanta, F. Estrada, P. Wils & S. Süsstrunk. *Salient region detection and segmentation*. In International Conference on Computer Vision Systems (ICVS), pages 66–75, 2008.

[Achanta 09a] R. Achanta, S. Hemami, F. Estrada & S. Süsstrunk. *Frequency-tuned Salient Region Detection*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 1597 – 1604, 2009.

[Achanta 09b] R. Achanta & S. Süsstrunk. *Saliency Detection for Content-aware Image Resizing*. In IEEE Intlernational Conference on Image Processing (ICIP), 2009.

[Achanta 10] R. Achanta & S. Süsstrunk. *Saliency detection using maximum symmetric surround*. In IEEE Intlernational Conference on Image Processing (ICIP), pages 2653–2656, 2010.

[Akman 10] O. Akman & P. Jonker. *Computing saliency map from spatial information in point cloud data*. In Advanced Concepts for Intelligent Vision Systems (ACIVS), pages 290–299, 2010.

[Aldoma 13] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. di Stefano & M. Vincze. *Multimodal Cue Integration through Hypotheses Verification for RGB-D Object Recognition and 6DOF Pose Estimation*. In IEEE International Conference on Robotics and Automation (ICRA), 2013.

[Alexe 12] B. Alexe, T. Deselaers & V. Ferrari. *Measuring the Objectness of Image Windows*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 34, no. 11, pages 2189–2202, 2012.

[Aloimonos 88] J. Aloimonos, I. Weiss & A. Bandyopadhyay. *Active Vision*. International Journal of Computer Vision (IJCV), vol. 1, no. 4, pages 333–356, 1988.

[Arbelaez 12] P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev & J. Malik. *Semantic Segmentation using Regions and Parts*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 3378–3385, 2012.

[Aziz 10] M. Z. Aziz & B. Mertsching. *Fast Depth Saliency from Stereo for Region-Based Artificial Visual Attention*. Advanced Concepts for Intelligent Vision Systems (ACIVS), vol. 6474, pages 367–378, 2010.

[Backer 00] G. Backer & B. Mertsching. *Integrating Time and Depth into the Attentional Control of an Active Vision System*, 2000.

[Begum 11] M. Begum & F. Karray. *Visual Attention for Robotic Cognition: A Survey*. IEEE Transactions on Autonomous Mental Development, vol. 3, no. 1, pages 92–105, 2011.

[Bergström 11] N. Bergström, M. Björkman & D. Kragic. *Generating Object Hypotheses in Natural Scenes Through Human-Robot Interaction*. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 827–833, 2011.

[Berner 11] A. Berner, M. Wand, N. J. Mitra, D. Mewes & H.-P. Seidel. *Shape Analysis with Subspace Symmetries*. Computer Graphics Forum, vol. 30, no. 2, pages 277–286, 2011.

[Biswas 12] J. Biswas & M. Veloso. *Depth Camera based Indoor Mobile Robot Localization and Navigation*, 2012.

[Björkman 10] M. Björkman & D. Kragic. *Active 3D Scene segmentation and Detection of Unknown Objects*. In IEEE International Conference on Robotics and Automation (ICRA), pages 3114–3120, 2010.

[Borji 11] A. Borji & L. Itti. *Scene Classification with a Sparse Set of Salient Regions*. In IEEE International Conference on Robotics and Automation (ICRA), pages 1902–1908, 2011.

[Borji 13] A. Borji & L. Itti. *State-of-the-art in Visual Attention Modeling*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 35, no. 1, pages 185–207, Jan 2013.

[Boykov 01] Y. Boykov & M.-P. Jolly. *Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images*. In IEEE International Conference on Computer Vision (ICCV), volume 1, pages 105–112 vol.1, 2001.

[Bruce 05] N. D. B. Bruce & J. K. Tsotsos. *An Attentional Framework for Stereo Vision*. Canadian Conference on Computer and Robot Vision, pages 88–95, 2005.

[Bruce 09] N. D. B. Bruce & J. K. Tsotsos. *Saliency, Attention, and Visual Search: An Information Theoretic Approach*. Journal of Vision (JOV), vol. 9, no. 3, 2009.

[Campbell 07] N. Campbell, G. Vogiatzis, C. Hernández & R. Cipolla. *Automatic 3D Object Segmentation in Multiple Views using Volumetric Graph-cuts*. In British Machine Vision Conference (BMVC), volume 28, pages 530–539, 2007.

[Chang 11] C.-C. Chang & C.-J. Lin. *LIBSVM: A Library for Support Vector Machines*. ACM Transactions on Intelligent Systems and Technology (ACM TIST), vol. 2, no. 3, pages 1–27, 2011.

[Cheng 11] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang & S.-M. Hu. *Global Contrast Based Salient Region Detection*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 409–416, 2011.

[Chertok 10] M. Chertok & Y. Keller. *Spectral Symmetry Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 32, no. 7, pages 1227–1238, 2010.

[Collet 11] A. Collet, S. S. Srinivasa & M. Hebert. *Structure Discovery in Multi-modal Data: A Region-based Approach*. In IEEE International Conference on Robotics and Automation (ICRA), pages 5695–5702, 2011.

[Comaniciu 02] D. Comaniciu & P. Meer. *Mean Shift: a Robust Approach toward Feature Space Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 24, no. 5, pages 603–619, 2002.

[Courty 03] N. Courty, E. Marchand & B. Arnaldi. *A New Application for Saliency Maps: Synthetic Vision of Autonomous Actors*. In IEEE Intlernational Conference on Image Processing (ICIP), pages 1065–1068, 2003.

[Cunha 11] J. Cunha, E. Pedrosa, C. Cruz, A. J. R. Neves & N. Lau. *Using a Depth Camera for Indoor Robot Localization and Navigation*. In Robotics Science and Systems (RSS): RGB-D Workshop on Advanced Reasoning with Depth Cameras, 2011.

[Desingh 13] K. Desingh, K. M. Krishna, D. Rajan & C. V. Jawahar. *Depth Really Matters: Improving Visual Salient Region Detection with Depth*. In British Machine Vision Conference (BMVC), 2013.

[Deville 08] B. Deville, G. Bologna, M. Vinckenbosch & T. Pun. *Guiding the Focus of Attention of Blind People with Visual Saliency*. In Workshop on Computer Vision Applications for the Visually Impaired, 2008.

[Einhäuser 08] W. Einhäuser, M. Spain & P. Perona. *Objects Predict Fixations Better than Early Saliency*. Journal of Vision (JOV), vol. 8, no. 14, pages 1–26, 2008.

[Felzenszwalb 04] P. F. Felzenszwalb & D. P. Huttenlocher. *Efficient Graph-Based Image Segmentation*. International Journal of Computer Vision (IJCV), vol. 59, no. 2, pages 167–181, 2004.

[Feng 11] J. Feng, Y. Wei, L. Tao, C. Zhang & J. Sun. *Salient Object Detection by Composition*. In International Conference on Computer Vision (ICCV), pages 1028–1035, 2011.

[Fischler 81] M. A. Fischler & R. C. Bolles. *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Communications of the ACM (CACM), vol. 24, pages 381–395, 1981.

[Frintrop 05a] S. Frintrop, G. Backer & E. Rome. *Goal-Directed Search with a Top-Down Modulated Computational Attention System*. In Annual Meeting of the German Association for Pattern Recognition (DAGM), pages 117–124, 2005.

[Frintrop 05b] S. Frintrop, E. Rome, A. Nüchter & H. Surmann. *A Bimodal Laser-based Attention System*. Computer Vision and Image Understanding (CVIU), vol. 100, pages 124–151, 2005.

[Frintrop 06a] S. Frintrop. VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, volume 3899 of *Lecture Notes in Computer Science*. Springer, 2006.

[Frintrop 06b] S. Frintrop, P. Jensfelt & H. I. Christensen. *Attentional Landmark Selection for Visual SLAM*. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2006.

[García 13] G. García, S. Frintrop & A. B. Cremers. *Attention-Based Detection of Unknown Objects in a Situated Vision Framework*. KI-Künstliche Intelligenz, vol. 27, no. 3, pages 267–272, 2013.

[Gautier 12] J. Gautier & O. Le Meur. *A Time-Dependent Saliency Model Combining Center and Depth Biases for 2D and 3D Viewing Conditions*. Cognitive Computation, vol. 4, no. 2, pages 141–156, 2012.

[Goferman 10] S. Goferman, L. Zelnik-Manor & A. Tal. *Context-aware Saliency Detection*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 2376–2383, 2010.

[González-Jiménez 13] J. González-Jiménez, J. R. Ruiz-Sarmiento & C. Galindo. *Improving 2D Reactive Navigators with Kinect*. In International Conference on Informatics in Control, Automation and Robotics (ICINCO), 2013.

[Gulshan 10] V. Gulshan, C. Rother, A. Criminisi, A. Blake & A. Zisserman. *Geodesic Star Convexity for Interactive Image Segmentation*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[Häkkinen 10] J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya & G. Nyman. *What Do People Look at When They Watch Stereoscopic Movies?* Electronic Imaging: Stereoscopic Displays and Applications XXI, SPIE, vol. 7524, pages 1–10, 2010.

[Harel 06] J. Harel, C. Koch & P. Perona. *Graph-based Visual Saliency*. Advances in Neural Information Processing Systems (NIPS), vol. 19, pages 545–552, 2006.

[Heidemann 04] G. Heidemann. *Focus-of-attention From Local Color Symmetries*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2004.

[Hou 07] X. Hou & L. Zhang. *Saliency Detection: A Spectral Residual Approach*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 1 –8, 2007.

[Hügli 05] H. Hügli, T. Jost & N. Ouerhani. *Model Performance for Visual Attention in Real 3D Color Scenes*. In IWINAC (2), pages 469–478, 2005.

[Huynh-Thu 11] Q. Huynh-Thu & L. Schiatti. *Examination of 3D Visual attention in Stereoscopic Video Content*. Proc. SPIE, vol. 7865, 2011.

[Itti 98] L. Itti, C. Koch & E. Niebur. *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 20, no. 11, pages 1254–1259, 1998.

[Itti 99] L. Itti & C. Koch. *A Comparison of Feature Combination Strategies for Saliency-based Visual Attention Systems*. In SPIE Conference on Human Vision and Electronic Imaging IV (HVEJ), volume 3644, pages 473–482, 1999.

[Itti 00] L. Itti & C. Koch. *A Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention*. Vision Research, vol. 40, no. 10-12, pages 1489–1506, 2000.

[Itti 01] L. Itti & C. Koch. *Computational Modelling of Visual Attention*. Nature Reviews Neuroscience, vol. 2, no. 3, pages 194–203, 2001.

[Itti 06] L. Itti & P. Baldi. *Bayesian Surprise Attracts Human Attention*. Advance Neural Information Processing Systems (NIPS), pages 547–554, 2006.

[Jansen 09] L. Jansen, S. Onat & P. König. *Influence of Disparity on Fixation and Saccades in Free Viewing of Natural Scenes*. Journal of Vision (JOV), vol. 9, no. 1, pages 1–19, 2009.

[Jost 05] T. Jost, N. Ouerhani, R. von Wartburg, R. Müri & H. Hügli. *Assessing the Contribution of Color in Visual Attention*. Computer Vision and Image Understanding (CVIU), vol. 100, no. 1-2, pages 107–123, 2005.

[Karpathy 13] A. Karpathy, S. Miller & L. Fei-Fei. *Object Discovery in 3D Scenes via Shape Analysis*. In IEEE International Conference on oRobtics and Automation (ICRA), pages 2088–2095, 2013.

[Katz 13] D. Katz, A. Venkatraman, M. Kazemi, A. J. Bagnell & A. Stentz. *Perceiving, Learning, and Exploiting Object Affordances for Autonomous Pile Manipulation*. In Robotics: Science and Systems Conference (RSS), June 2013.

[Khaustova 13] D. Khaustova, J. Fournier, E. Wyckens & O. Le Meur. *How Visual Attention is Modified by Disparities and Textures Changes?* In Proceedings SPIE, volume 8651, 2013.

[Ko 06] B. C. Ko & J.-Y. Nam. *Object-of-interest Image Segmentation Based on Human Attention and Semantic Region Clustering*. Journal of the Optical Society of America A (JOSA A), vol. 23, no. 10, pages 2462–2470, October 2006.

[Koch 85] C. Koch & Ullman. S. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*. Human Neurobiology, vol. 4, pages 219–227, 1985.

[Kootstra 08] G. Kootstra, A. Nederveen & B. de Boer. *Paying Attention to Symmetry*. In British Machine Vision Conference (BMVC), pages 1115–1125, 2008.

[Kootstra 10a] G. Kootstra, N. Bergström & D. Kragic. *Fast and Automatic Detection and Segmentation of Unknown Objects*. In IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS), pages 442–447, 2010.

[Kootstra 10b] G. Kootstra, N. Bergström & D. Kragic. *Using Symmetry to Select Fixation Points for Segmentation*. In International Conference on Pattern Recognition (ICPR), pages 3894–3897, aug. 2010.

[Lai 12a] K. Lai, L. Bo, X. Ren & D. Fox. *Detection-based Object Labeling in 3D Scene*. In IEEE International Conference on Robotics and Automation (ICRA), 2012.

[Lai 12b] K. Lai, L. Bo, X. Ren & D. Fox. *RGB-D Object Dataset*, Oct. 2012.

[Lang 12a] C. Lang, G. Liu, J. Yu & S. Yan. *Saliency Detection by Multitask Sparsity Pursuit*. IEEE Transactions on Image Processing, vol. 21, no. 3, pages 1327–1338, 2012.

[Lang 12b] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli & S. Yan. *Depth Matters: Influence of Depth Cues on Visual Saliency*. In European Conference on Computer Vision (ECCV), pages 101–115, 2012.

[Le Meur 06] O. Le Meur, P. Le Callet, D. Barba & D. Thoreau. *A Coherent Computational Approach to Model Bottom-Up Visual Attention*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 28, no. 5, pages 802–817, 2006.

[Lee 99] D. K. Lee, L. Itti, C. Koch & J. Braun. *Attention Activates Winner-Take-All Competition Among Visual Filters*. Nature Neuroscience, vol. 2, no. 4, pages 375–381, 1999.

[Lee 05] C. H. Lee, A. Varshney & D. W. Jacobs. *Mesh Saliency*. In ACM SIGGRAPH, pages 659–666, 2005.

[Li 13] J. Li, M. D. Levine, X. An, X. Xu & H. He. *Visual Saliency Based on Scale-Space Analysis in the Frequency Domain*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 35, no. 4, pages 996–1010, April 2013.

[Liu 10] Y. Liu, L. K. Cormack & A. C. Bovik. *Natural Scene Statistics at Stereo Fixations*. In Symposium on Eye-Tracking Research and Applications, pages 161–164, 2010.

[Loy 03] G. Loy & A. Zelinsky. *Fast Radial Symmetry for Detecting Points of Interest*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 25, no. 8, pages 959–973, 2003.

[Mahadevan 12] V. Mahadevan & N. Vasconcelos. *On the Connections between Saliency and Tracking*. In Advance Neural Information Processing Systems (NIPS), pages 1673–1681, 2012.

[Maki 96] A. Maki, P. Nordlund & J. O. Eklundh. *A Computational Model of Depth-based Attention*. In International Conference on Pattern Recognition (ICPR), pages 734–739, 1996.

[Malik 01] J. Malik, S. Belongie, T. Leung & J. Shi. *Contour and Texture Analysis for Image Segmentation*. International Journal of Computer Vision (IJCV), vol. 43, no. 1, pages 7–27, 2001.

[Manén 13] S. Manén, M. Guillaumin & L. Van Gool. *Prime Object Proposals with Randomized Prim's Algorithm*. In International Conference on Computer Vision (ICCV), 2013.

[Martin 04] D. Martin, C. Fowlkes & J. Malik. *Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 26, no. 5, pages 530–549, May 2004.

[Minovic 93] P. Minovic, S. Ishikawa & K. Kato. *Symmetry Identification of a 3D Object Represented by Octree*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 15, pages 507–514, 1993.

[Mishra 09] A. Mishra & Y. Aloimonos. *Active Segmentation With Fixation*. In International Conference on Computer Vision (ICCV), 2009.

[Mishra 11] A. Mishra & Y. Aloimonos. *Visual Segmentation of Simple Objects for Robots*. In Robotics: Science and Systems Conference (RSS), 2011.

[Mitra 12] N. J. Mitra, M. Pauly, M. Wand & D. Ceylan. *Symmetry in 3D Geometry: Extraction and Applications*. In EUROGRAPHICS: State-of-the-art Report, 2012.

[Mörwald 13] T. Mörwald, A. Richtsfeld, J. Prankl, M. Zillich & M. Vincze. *Geometric Data Abstraction Using B-splines for Range Image Segmentation*. In IEEE International Conference on Robotics and Automation (ICRA), pages 148–153, 2013.

[Navalpakkam 06] V. Navalpakkam & L. Itti. *An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 2049 – 2056, 2006.

[Newcombe 11] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges & A. Fitzgibbon. *KinectFusion: Real-time Dense Surface Mapping and Tracking*. In IEEE International Symposium on Mixed and Augmented Reality, pages 127–136, 2011.

[Niu 12] Y. Niu, Y Geng, X. Li & Liu F. *Leveraging stereopsis for saliency analysis*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 454–461, 2012.

[Ouerhani 00] N. Ouerhani & H. Hügli. *Computing Visual Attention from Scene Depth*. In International Conference on Pattern Recognition (ICPR), pages 375–378, 2000.

[Ouerhani 01] N. Ouerhani, N. Archip, H. Hügli & P.-J. Erard. *Visual Attention Guided Seed Selection for Color Image Segmentation*. In International Conference on Computer Analysis of Images and Patterns, pages 630–637, 2001.

[Paletta 13] L. Paletta, K. Santner, G. Fritz, A. Hofmann, G. Lodron, G. Thallinger & H. Mayer. *FACTS - A Computer Vision System for 3D Recovery and Semantic Mapping of Human Factors*. In International Conference on Computer Vision Systems (ICVS), pages 62–72, 2013.

[Pirri 11] F. Pirri, M. Pizzoli & A. Rudi. *A General Method for the Point of Regard Estimation in 3D Space*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 921–928, 2011.

[Posner 84] M. I. Posner & Y. Cohen. *Components of Visual Orienting*. Attention and Performance X, vol. 32, pages 531–556, 1984.

[Potapova 11] E. Potapova, M. Zillich & Vinczem M. *Learning What Matters: Combining Probabilistic Models of 2D and 3D Saliency Cues*. In International Conference on Computer Vision Systems (ICVS), 2011.

[Potapova 12a] E. Potapova, M. Zillich & M. Vincze. *Attention-driven Segmentation of Cluttered 3D Scenes*. In International Conference on Pattern Recognition (ICPR), pages 3610–3613, 2012.

[Potapova 12b] E. Potapova, M. Zillich & M. Vincze. *Local 3D Symmetry for Visual Saliency in 2.5D Point Clouds*. In Asian Conference on Computer Vision (ACCV, pages 434–445, 2012.

[Potapova 14a] E. Potapova, , A. Richtsfeld, M. Zillich & M. Vincze. *Incremental Attention-driven Object Segmentation*. In IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS), 2014.

[Potapova 14b] E. Potapova, K. M. Varadarajan, A. Richtsfeld, M. Zillich & M. Vincze. *Attention-Driven Object Detection and Segmentation of Cluttered Table Scenes Using 2.5D Symmetry*. In IEEE International Conference on Robotics and Automation (ICRA), pages 3610–3613, 2014.

[Rahtu 11] E. Rahtu, J. Kannala & M. B. Blaschko. *Learning a Category Independent Object Detection Cascade*. In International Conference on Computer Vision (ICCV), pages 1052–1059, 2011.

[Ramasamy 09] C. Ramasamy, D. H. House, A. T. Duchowski & B. Daugherty. *Using Eye Tracking to Analyze Stereoscopic Filmmaking*. In SIGGRAPH: Posters, pages 28:1–28:1, 2009.

[Reisfeld 95] D. Reisfeld, H. Wolfson & Y. Yeshurun. *Context Free Attentional Operators: the Generalized Symmetry Transform*. International Journal of Computer Vision (IJCV), vol. 14, pages 119–130, 1995.

[Richtsfeld 12] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich & M. Vincze. *Segmentation of Unknown Objects in Indoor Environments*. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4791–4796, 2012.

[Richtsfeld 13] A. Richtsfeld, T. Mörwald, J. Prankl & M. Zillich. *The Object Segmentation Database (OSD)*, Mar. 2013.

[Rother 04] C. Rother, V. Kolmogorov & A. Blake. *GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts*. ACM Transactions on Graphics, vol. 23, no. 3, pages 309–314, 2004.

[Rusu 09] R. B. Rusu. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments*. PhD thesis, Computer Science Department, Technische Universität München, Germany, October 2009.

[Scholl 01] B. J. Scholl. *Objects and Attention: The State-of-the-art*. Cognition, vol. 80, no. 1-2, pages 1–46, 2001.

[Sedlacek 09] D. Sedlacek & J. Zara. *Graph-Cut Based Point Cloud Segmentation for Polygonal Reconstruction*. In International Conference on Computer Vision Systems (ICVS), pages 218–227, 2009.

[Seo 09] H. J. Seo & P. Milanfar. *Static and Space-time Visual Saliency Detection by Self-Resemblance*. Journal of Vision (JOV), vol. 9, page 15, 2009.

[Shi 00] J. Shi & J. Malik. *Normalized Cuts and Image Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 8, pages 888–905, 2000.

[Shotton 11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman & A. Blake. *Real-time Human Pose Recognition in Parts from Single Depth Images*. In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 1297–1304, 2011.

[Strom 10] J. Strom, A. Richardson & E. Olson. *Graph-Based Segmentation for Colored 3D Laser Point Clouds*. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2131–2136, 2010.

[Sun 97] C. Sun & J. Sherrah. *3D Symmetry Detection Using The Extended Gaussian Image*. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 19, pages 164–168, 1997.

[Treisman 80] A. M. Treisman & G. Gelade. *A Feature-Integration Theory of Attention*. Cognitive Psychology, vol. 12, no. 1, pages 97–136, 1980.

[Triebel 10] R. Triebel, J. Shin & R. Siegwart. *Segmentation and Unsupervised Part-based Discovery of Repetitive Objects*. In Robotics: Science and Systems Conference (RSS), 2010.

[Tsotsos 90] J. K. Tsotsos. *Analyzing Vision at the Complexity Level*. Behavioral and Brain Sciences, vol. 13, no. 3, pages 423–469, 1990.

[Tsotsos 07] J. K. Tsotsos & K. Shubina. *Attention and Visual Search : Active Robotic Vision Systems that Search*. In International Conference on Computer Vision Systems (ICVS), 2007.

[Ückermann 12] A. Ückermann, R. Haschke & H. Ritter. *Real-Time 3D Segmentation of Cluttered Scenes for Robot Grasping*. In IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS), pages 1734–1740, 2012.

[Varadarajan 12] K. M. Varadarajan & Vinczem M. *MRF guided Anisotropic Depth Diffusion for Kinect Range Image Enhancement*. In Asian Conference on Computer Vision (ACCV, 2012.

[Viola 04] P. Viola & M. J. Jones. *Robust Real-Time Face Detection*. International Journal of Computer Vision, vol. 57, no. 2, pages 137–154, May 2004.

[Vishwanath 04] D. Vishwanath & E. Kowler. *Saccadic Localization in the Presence of Cues to Three-dimensional Shape*. Journal of Vision (JOV), vol. 4, pages 445–458, 2004.

[Walther 04] D. Walther, U. Rutishauser, C. Koch & Perona. P. *On the Usefulness of Attention for Object Rrecognition*. In Workshop on Attention and Performance in Computational Vision at ECCV, pages 96–103, 2004.

[Walther 06] D. Walther & C. Koch. *Modeling Attention to Salient Proto-Objects*. Neural Networks, vol. 19, no. 9, pages 1395–1407, 2006.

[Wan 12] J. Wan, T. Xia, S. Tang & J. Li. *Robust Range Image Segmentation Based on Coplanarity of Superpixels*. In International Conference on Pattern Recognition (ICPR), pages 3618–3621, 2012.

[Wang 12] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel & M. P. Da Silva. *Study of Depth Bias of Observers in Free Viewing of Still Stereoscopic Synthetic Stimuli*. Journal Of Eye Movement Research, vol. 5, no. 5, 2012.

[Wang 13] J. Wang, M. M. Da Silva, P. Le Callet & V. Ricordel. *Computational Model of Stereoscopic 3D Visual Saliency*. IEEE Transactions on Image Processing, vol. 22, no. 6, pages 2151–2165, 2013.

[Wexler 08] M. Wexler & N. Ouarti. *Depth Affects Where We Look*. Current Biology, vol. 18, pages 1872–1876, 2008.

[Wismeijer 10] D. A. Wismeijer, C. J. Erkelens, R. van Ee & M. Wexler. *Depth Cue Combination in sSpontaneous Eye Movements*. Journal of Visision (JOV), vol. 10, no. 6, page 25, 2010.

[Wolfe 89] J. M. Wolfe, K. R. Cave & S. L. Franzel. *Guided Search: An Alternative to the Feature Integration Model for Visual Search*. Journal of Experimental Psychology: Human Perception and Performance, vol. 15, no. 3, pages 419–433, 1989.

[Wolfe 03] J. Wolfe, S. Butcher, C. Lee & Hyle M. *Changing Your Mind: On the Contributions of Top-down and Bottom-up Guidance in Visual Search for Feature Singletons*. Journal of Experimental Psychology, vol. 29, no. 2, pages 483–502, 2003.

[Y.-M. 07] Jangm Y.-M., S.-W. Ban & M. Lee. *Stereo Saliency Map Considering Affective Factors in a Dynamic Environment*. In ICONIP (2), pages 1055–1064, 2007.

[Yarbus 67] A. Yarbus. *Eye Movements and Vision*. New York, USA: Plenum, 1967.

[Yu 09] S. X. Yu & D. A. Lisin. *Image Compression Based on Visual Saliency at Individual Scales*. In International Symposium on Visual Computing (ISVC), pages 157–166, 2009.

[Zhai 06] Y. Zhai. *Visual Attention Detection in Video Sequences Using Spatiotemporal Cues*. ACM Multimedia, 2006.

[Zhang 08] L. Zhang, M. Tong, T. Marks, H. Shan & Cottrell G. *SUN: A Bayesian Framework for Saliency Using Natural Statistics*. Journal of Vision (JOV), vol. 8, no. 7, pages 1–20, 2008.

[Zhang 10] Y. Zhang, G. Jiang, M. Yu & K. Chen. *Stereoscopic Visual Attention Model for 3D Video*. In Conference on Multimedia Modeling, pages 314–324, 2010.

[Zhou 00] H. Zhou, H. S. Friedman & R. Von Der Heydt. *Coding of Border Ownership in Monkey Visual Cortex*. Journal of Neuroscience, vol. 20, pages 6594–6611, 2000.