Paladyn. Journal of Behavioral Robotics manuscript No.
(will be inserted by the editor)

Towards Attentive Robots

Simone Frintrop[1]

Received: date / Accepted: date

Abstract This paper introduces *Attentive Robots*: robots that attend to the parts of their sensory input that are currently of most potential interest. The concept of selecting the most promising parts is adopted from human perception where selective attention allocates the brain resources to the most interesting parts of the sensory input. We give an overview of current approaches to integrate computational attention into robotic systems, with a focus on biologically-inspired visual attention methods. Example applications range from localization with salient landmarks over object manipulation to the design of social robots. A brief outlook gives an impression of how future ways to obtain attentive robots might look like.

Keywords

1 Introduction

Imagine you bought a new home robot, Dobby, at some point in the future. Dobby is supposed to do most of the housework while you are at work or are meeting with friends. It shall receive and unpack the groceries that come from the supermarket, do the laundry, and tidy up the mess that the kids made when playing in the living room. At every moment, Dobby has to process a large amount of sensory input and the possibilities of what to do first easily become overwhelming. Since robots have limited processing power as well as physical limitations such as a limited number of sensors, arms, etc., a selection mechanism that determines where to concentrate the resources is of high interest. In humans, the mechanism that determines which part of the sensory input is currently most promising is called selective attention [Pashler, 1997]. Accordingly, we call robots that attend to the most promising part of their sensor data "Attentive Robots" (cf. Fig. 1).

The term "attention" is used in many contexts and many definitions exist. It is a term of common language (William James: "Everyone knows what attention is..."

S. Frintrop
Institute of Computer Science III, Rheinische Friedrich-Wilhelms-Universität,
53117 Bonn, Germany.
E-mail: frintrop@iai.uni-bonn.de

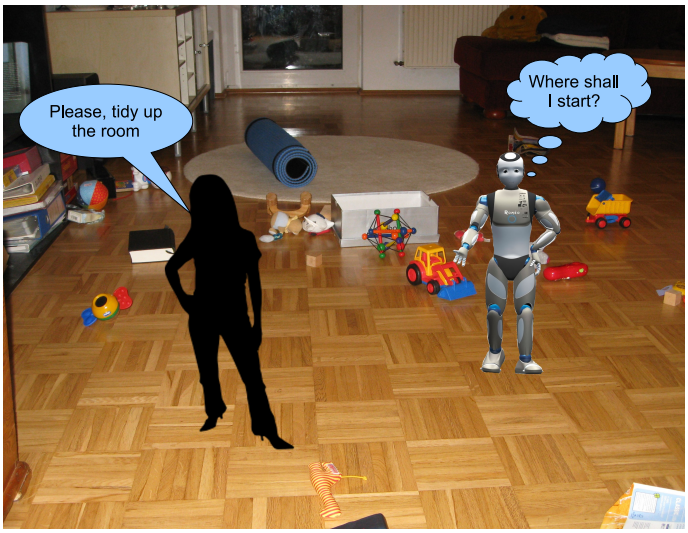
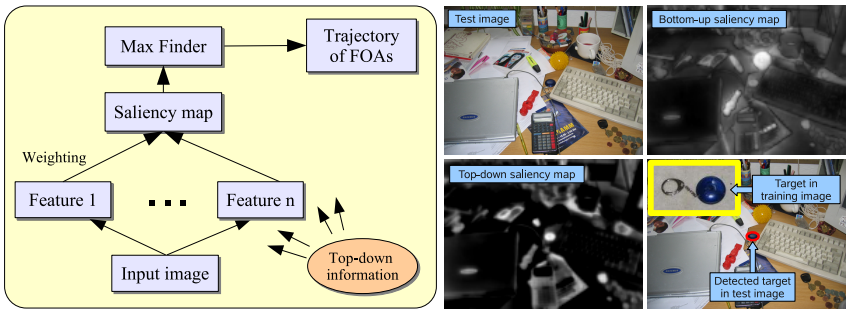


Fig. 1 The scene visualizes the concept of an attentive robot: to tidy up the room, the robot has to investigate the scene and therefore attend to the objects on the floor, one object at a time. An attention module endows it with the capability to focus on regions of most potential interest. This enables efficient processing and prioritizes the robot's actions.

[James, 1890]), it is an active research area in psychophysics since many decades, and it is frequently used in machine vision and robotics to refer to mechanisms that focus further processing on regions of interest. The latter perspective of attention is very broad, in principle, any pre-processing method of sensor data could be called attentional since it focuses further processing on parts of the data. We believe that closely mimicking the human system has the advantage that it results in human-like behavior, which is beneficial for systems that should interact with humans in a natural and intuitive manner. Therefore, in this article, we focus on methods that are based on concepts of human perception. Following this direction, one of the best fitting definitions of attention comes from Wikipedia: "Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things" [1].

While the concept of attention exists for all senses, most research focuses on the visual part of attention. This is true both for human visual attention, due to the fact that vision is the most important sense in humans, and for computational attention system. Thus, with a few exceptions the approaches mentioned in this article focus on analyzing visual data.

During the last decade, attentional modules for autonomous robots have significantly gained in popularity. The reasons are two-fold. First, adequate computational resources are now available to compute the focus of attention in real-time [Frintrop et al., 2007, Xu et al., 2009] and the methods are robust enough to deal with real-world conditions. Second, basic techniques such as localization and collision avoidance have reached a quite mature level and interest has moved on to higher level tasks and challenges. The more complex a system becomes, the more urgent is the need for optimizing the visual processing. The high level of interest



(a) (b)

Fig. 2 (a) General structure of most visual attention systems. (b) Example of saliency computation with the attention system VOCUS: bottom-up exploration (top-right) and top-down search for the target "key fob" (bottom).

in such capabilities has led to a large number of EU projects on Cognitive Systems during the last decade. Many of the robots developed in these projects have an attentional module, e.g., in the projects MACS, PACO-PLUS, RobotCub, and GRASP.

In this paper, we will give an overview of the current state of the art in computational attention systems for autonomous robots. While being far from an exhaustive overview, we aim to give the reader an impression of what attention systems can do for cognitive robots and which directions already exist. Finally, we briefly discuss possible future ways to lead us closer to the dream of attentive robots.

2 The Basic Structure of Computational Attention Systems

Most biologically inspired attention systems have a similar structure, which is depicted in Fig. 2 (a). This structure is originally adapted from psychological theories like the Feature Integration Theory [Treisman and Gelade, 1980] and the Guided Search model [Wolfe, 1994]. The main idea is to compute several feature channels such as intensity, color, orientation or motion, in parallel and to fuse their conspicuities in a saliency map. This map is a gray-level image with pixel brightness proportional to the saliency (cf. Fig. 2 (b), top right). This approach is adopted from the parallel processing of different features in the human brain; some brain areas are mainly involved in processing color while others concentrate on motion processing and so on [Palmer, 1999]. If top-down information is available, e.g., prior knowledge on the context, the task, the searched object, etc., it can be used to influence the processing. An example of top-down attention is shown in Fig. 2 (b), bottom row. Here, knowledge about the target object "key fob" is fed into the attention system as a feature descriptor, resulting in a top-down saliency map (details in [Frintrop, 2006]).

The feature computations are usually based on contrast computations with *center-surround filters*. Such filters are inspired by cells in the human visual system (e.g. ganglion cells in the retina) that compute the contrast of a center and

a surround region [Palmer, 1999]. Computationally, they are usually modeled by Difference-of-Gaussian or Gabor filters. The feature channels most frequently implemented in computational attention systems are intensity, color, orientation, and motion.

One of the most important capabilities of attention systems is their ability to detect regions that differ from the rest of the image, a property that makes an object "salient". That means, the saliency of an object depends on the context. A red ball on grass is salient, while it is not salient among other red balls. Therefore, attention systems usually weight the feature maps according to the uniqueness of the feature. Feature maps with much activation obtain a low weight while those with few strong activation peaks obtain a high weight (details in [Frintrop, 2006]).

After obtaining a saliency map, the maxima in this map denote the regions that are investigated by the focus of attention (FOA) in the order of decreasing saliency. This trajectory of FOAs imitates human eye movements. Output of a computational attention system is either the saliency map itself or a trajectory of focused regions.

While most attention systems share this general structure, there are different ways of implementing the details. One of the best known computational models is the iNVT from the group around Itti [Itti et al., 1998]. In our group, we have developed the VOCUS model [Frintrop, 2006, 2011], that has adopted and extended several ideas from the iNVT. It is real-time capable and has a top-down mode to search for objects. Tsotsos and his group have developed the selective tuning model. A full description of the model and an overview of attention theories are available in his recent book [Tsotsos, 2011]. During the last years, some approaches came up that use information-theoretic concepts to determine visual saliency [Bruce and Tsotsos, 2009, Gao et al., 2009]. A survey on the cognitive foundations and state of the art of computational attention systems can be found in [Frintrop et al., 2010], an introduction to the topic for students and people new to the field is available in [Frintrop, 2011].

As mentioned in the introduction, this paper focuses on approaches that are based on concepts of human perception and share the above structure. However, it is worth noting that numerous approaches exist that compute saliency in ways that are less or not at all biologically-motivated. For example, Hou and Zhang [2007] compute the spectral residual of an image in the frequency domain and Gould et al. [2007] and Liu et al. [2009] learn an optimal feature combination with machine learning techniques.

3 Attentive Robots: The State of the Art

A future attentive robot is supposed to use attention for many tasks, on different levels of abstraction. If Dobby shall tidy up the room, it must focus on objects at unusual places and has to know where each object belongs. If it shall bring you the salt shaker, it should focus on the cupboard where the shaker is usually stored and should concentrate on features fitting to the appearance of the shaker. If you give Dobby an order, it has to interpret your gestures, facial expressions, and voice, e.g., it should follow your gaze and your pointing finger.

The tasks of the robot that involve visual attention might be classified roughly into three categories. The first, most low-level category, uses attention to detect

salient landmarks that can be used for localization and scene recognition (sec. 3.1). The second, mid-level category considers attention as a front-end for object recognition (sec. 3.2). In the third, highest-level category, attention is used in a human-like way to guide the actions of an autonomous system like a robot, i.e., to guide object manipulation or human-robot interaction (sec. 3.3).

3.1 Salient Landmarks

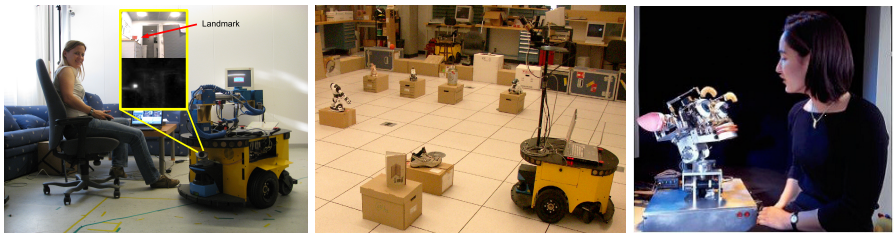
A basic capability of autonomous mobile robots is to localize themselves in their environment. Based on a known map of the surrounding, the robot has to determine its position in this map by interpreting its sensor data. When based on visual data, this is done by detecting visual landmarks with a known position. A visual landmark can be anything that the robot can see: a blob on the wall, a corner of an object, the edge of a door, or the door itself. One of the primary requirements of a visual landmark is that it should be redetectable under changing illumination conditions and from new viewpoints. It should also be possible to compute it quickly and to store it without much effort. Therefore complex object descriptions are seldom used. Salient landmarks are excellent candidates for localization since they have a high uniqueness. This makes them easy to redetect and diminishes the risk of confusing them with other landmarks. This also enables a landmark detection algorithm to concentrate on a sparse set of landmarks which reduces computation complexity. We have shown that the repeatability of salient regions in different scenes is significantly higher than the repeatability of standard detectors [Frintrop, 2008].

An early project that used salient landmarks for **localization** was the ARK project [Nickerson et al., 1998]. It relied on hand-coded maps, including the locations of known static obstacles as well as the locations of natural visual landmarks. Siagian and Itti [2009] presented an approach for **scene classification** and global localization based on salient landmarks. Additionally to the landmarks, the authors use the "gist" of the scene, a feature vector which captures the appearance of the scene, to obtain a coarse localization hypothesis.

In the above examples, a map of the environment is initially known. A more difficult task is **simultaneous localization and mapping (SLAM)** in which a robot has to build a map and localize itself inside it at the same time. We investigated the combination of visual attention and SLAM in [Frintrop and Jensfelt, 2008]. Salient regions are detected with the attention system VOCUS, tracked over several frames to obtain a 3D position of the landmarks, and matched to database entries of all previously seen landmarks. This enables the robot to detect if it closed a loop (see Fig. 3 (a)). Active camera control facilitated the redetection of landmarks.

3.2 Supporting Object Detection and Recognition

In addition to navigation, object detection and recognition are important tasks for autonomous robots, especially for manipulating objects. The terms object detection, object localization, object recognition, and classification are closely related and often used interchangeably. Let us therefore clarify our understanding of the



(a) Dumbo: Attentive visual SLAM (b) Curious George: Attentive object detection (c) Kismet: An Attentive, Social Robot

Fig. 3 Three application scenarios for visual attention systems: (a) Simultaneous localization and mapping (SLAM): robot Dumbo corrects its position estimate by redetecting a landmark which it has seen before. Landmark detection is done with the attention system VOCUS. The yellow rectangle shows the view of the robot: an image with a landmark and the corresponding saliency map (Fig. from http://www.iai.uni-bonn.de/∼frintrop/research.html) (b) Curious George: attention regions are detected in a peripheral camera image and investigated in detail by a foveal camera (Fig. from Forssén et al. [2008]). (c) Kismet is a social robot that interacts with people. Its gaze is controlled by a visual attention system (Fig. from [Breazeal, 2000] © Sam Ogden).

terms. Object detection or localization tackles the problem of localizing objects in images, e.g., by providing a bounding box around the object. Usually, the object is comparably small in the scene which makes the task challenging. The object to find might be a specific object (my favorite cup) or, as in the PASCAL VOC object detection challenge [Everingham et al., 2010], any instance of a certain class (any cup). In psychological literature on visual perception, the task to find an object is usually called visual search. A candidate to solve the visual search problem is top-down tuned visual attention [Frintrop, 2006]. Localizing an object often involves recognizing it, but may also be restricted to providing location candidates that are classified in a second step. The detection of any instance of an object class in cluttered images is still largely unsolved. In the latest PASCAL VOC 2010 challenge [Everingham et al., 2010], the best methods for object detection achieved only an average precision between 13% (potted plants) and 58.4% (aeroplanes). Sometimes, object detection refers also to the task to find anything in the scene that is an object (also called general object detection). Bottom-up attention is a perfect candidate for this kind of task since it does not require any prior knowledge on the objects.

Object classification deals with finding all instances of a certain class in a scene, e.g. faces. It is usually applied to pre-segmented objects or it uses a sliding windows approach, in which subregions of the images are successively investigated by the classifier. The term recognition is mostly used for the recognition of instances but is sometimes also used as synonym for classification.

Visual attention methods are of special interest for all tasks in which the object is comparably small in the image, as in object detection and localization or in classification on non pre-segmented images. These tasks become considerably easier if an attentional mechanism first focuses the processing on regions of potential interest. Thus is because of two reasons. First, this reduces the search space and results in reduction in computational complexity. Second, most recognition and

classification methods work best if the object occupies a dominant portion of the image.

Several approaches have been proposed to use visual attention as preprocessing step for **classification** or **object detection**. Miau et al. [2001] present a biologically motivated approach that combines an attentional front-end with the biologically motivated object recognition system HMAX. The experiments are restricted to recognize simple artificial objects like circles or rectangles. Alternatively, the authors have used a support vector machine to detect pedestrians in natural images. Walther [2006] combine their Saliency Toolbox, a Matlab implementation of the iNVT, with an object recognizer based on SIFT features and show that the recognition results are improved by the attentional front-end. Vogel and de Freitas [2008] combine the iNVT with a classifier to perform gaze planning in complex scenes. In the above mentioned approaches, the attentional part is separated from the object recognition; both systems work independently. In human perception, these processes are strongly intertwined. Accordingly, Walther and Koch [2007] suggest a unifying framework for object recognition and attention. It is based on the HMAX model and modulates the activity by spatial and feature modulation functions which suppress or enhance locations or features due to spatial attention.

While the above approaches are not applied in a robotics context, some groups have recently integrated attentive object detection on real robots. Two approaches that determine regions of interest with visual attention in a peripheral vision system, focus on these regions with a foveal vision system, and investigates these high-resolution images with an object recognition method are presented in [Gould et al., 2007] and [Meger et al., 2008]. The robot in the latter approach, curious George, placed first in the robot league of the Semantic Robot Vision Challenge (SRVC)[1] both in 2007 and 2008, and first in the software league for 2009 (see also Fig. 3 (b)).

All of these systems rely only on bottom-up information and therefore on the assumption that the objects of interest are sufficiently salient by themselves. For some object classes like traffic signs or toys, which are intentionally designed salient, this works quite well; for other applications, top-down information is needed to enable the system to focus on the desired objects. A combination of a top-down modulated computational attention system with a classifier is presented by Mitri et al. [2005]. Here, the attention system VOCUS generates object hypotheses which are verified or falsified by a classifier. For the application of ball detection in the robot soccer scenario RoboCup, the amount of false detections is reduced significantly. Recently, Xu et al. [2010] have used visual bottom-up and top-down attention to detect objects with the Autonomous City Explorer (ACE) robot.

Some groups have used attentive object detection to support **object manipulation** on robots or robot arms. One of the earliest works on this topic was presented by Bollmann et al. [1999]: a Pioneer1 robot used the active vision system NAVIS to play at dominoes. The group around Tsotsos is working on a smart wheelchair to support disabled children [Tsotsos et al., 1998, Rotenstein et al., 2007]. The wheelchair has a display as easily accessible user interface which shows pictures of places and toys. Once a task like "go to table, point to toy" is selected, the system drives to the selected location and searches for the specified toy, us-

[1] http://www.semantic-robot-vision-challenge.org/

ing mechanisms based on a visual attention system. Rasolzadeh et al. [2010] use bottom-up and top-down attention to control a KUKA arm for detecting, recognizing, and grasping objects on a table. In [Björkman and Kragic, 2010] and [Johnson-Roberson et al., 2010] the FOAs from the same attention system were used as seeds for 3D segmentation of objects from stereo data.

3.3 Guiding Robot Action

A robot which has to act in a complex world faces the same problems as a human: it has to decide what to do next. Such decisions include where to go (drive), what to look at, what to grasp, and who to interact with. Thus, even if computational power would allow it to find all correspondences, to recognize all objects in an image, and process everything of interest, it would still be necessary to filter out the relevant information to determine the next action [Mehta et al., 2000, Loach et al., 2008]. This decision is based first, on the current sensor input and second, on the internal state, for example the current tasks and goals.

A field in which the decision about the next action is intrinsically based on visual data is **active vision**, i.e., the problem of where to look next [Bajcsy, 1985]. It deals with controlling "the geometric parameters of the sensory apparatus ... in order to improve the quality of the perceptual results" [Aloimonos et al., 1988]. Thus, it directs the camera to regions of potential interest as the human visual system directs the gaze, the head, and even the body of a person. Since visual attention triggers this control in humans, it is also an intuitive candidate for the active vision problem on machines. In Sec. 4, we discuss the relation of visual attention and the active vision problem in more detail, let us here focus on approaches that have used visual attention to perform active vision control.

One of the first active vision systems that integrated visual attention was presented by Clark and Ferrier [1988]. They describe how to steer a binocular robotic head with visual attention and perform simple experiments to fixate and track the most salient region in artificial scenes composed of geometric shapes. Vijayakumar et al. [2001] present an attention system which is used to guide the gaze of a humanoid robot. The authors consider only one feature, visual flow, which enables the system to attend to moving objects. To simulate the different resolutions of the human eye, two cameras per eye are used: one wide-angle camera for peripheral vision and one narrow-angle camera for foveal vision. In more recent work, the humanoid robot iCub bases its decisions to move eyes and neck on visual and acoustic saliency maps [Ruesch et al., 2008]. Additionally, all the object manipulation approaches of the previous section include active vision to focus on the detected objects.

In the future, we want to interact with robots as naturally and intuitively as possible. Studies in the field of **human-robot interaction** have shown that humans treat robots like people [Nass and Moon, 2000, Fong et al., 2003]. The more human-like the robot acts, the easier the communication with a human. An essential part for purposefully interacting with humans is to generate a joint focus of attention. A computational attention system similar to the human one can help a robot to focus on the same region as a human. According to this, Breazeal [1999] introduced the social robot Kismet that interacts with humans in a natural and intuitive way. Its gaze is controlled by a visual attention system (see Fig. 3 (c)).

For humans, following pointing gestures of other humans is an important ability to jointly focus their attention on objects of interest. Approaches to endow robots with a similar capability were proposed by Heidemann et al. [2004] and by Schauerte et al. [2010]. They analyze the direction of a pointing finger and fuse this top-down information with the bottom-up saliency of objects. A robot that learns visual scene exploration by imitating human gaze shifts is presented by Belardinelli [2008]. Nagai [2009] developed an action learning model based on spatial and temporal continuity of bottom-up features.

Finally, Muhl et al. [2007] presented an interesting sociological study in which the interaction of a human with a robot simulation is investigated. A robot face on a screen attends to objects, shown by a human, with help of a visual attention system. If the robot was artificially diverted and directed its gaze away from the object, humans tried to reobtain the robots attention by waving hands, making noise, or approaching to the robot. This shows that people established a communicative space with the robot and accepted it as a social partner.

4 Discussion and Outlook

This paper gives an overview of the state of the art in the field of computational visual attention for mobile robots. Several fields are related to the computation of attention and we will briefly discuss the similarities and differences to some of them. First, the computation of visual saliency clearly has some similarities to the computation of interest points or regions. Both approaches compute a local contrast within some feature dimension, some use even the same methods, e.g., Difference of Gaussians [Lowe, 2004]. The main difference is that standard interest points are local methods that are only influenced by a small local neighborhood, while salient regions are defined by the context. They "stick out of the scene" and thus, the whole scene or at least a large neighborhood influences the saliency of a region. Both methods are usually computed on several scales, but interest points use smaller scales than visual saliency, leading to smaller regions that influence the point and usually to a large amount of points per image (usually several hundreds or even thousands). This is reasonable and useful for tasks such as object recognition or image registration but less so for controlling the camera. Salient regions on the other hand are usually computed on larger scales to consider context information. Additionally, the uniqueness of features is computed that takes into consideration the global (or large local) surround of the region and is usually implemented as a non-linear weighting on top of the center-surround feature computations [Itti et al., 1998, Frintrop, 2006]. This method favors regions that occur seldom in the scene, an essential aspect of visual saliency. Additionally, classical interest points are usually restricted to one feature dimension (e.g. intensity or color contrast), while visual attention systems integrate the results from several feature channels. Finally, a strength of visual attention systems is that top-down information can be integrated easily into the system.

As mentioned in Sec. 3, there is also a strong relation of visual attention to the active vision problem. To distinguish the two, it is worth clarifying that visual attention is a method that can be applied to different problems while the active vision problem is a problem that looks for a method to solve it. Visual attention claims to focus the processing resources to regions of most potential interest. That

makes it a perfect candidate to solve the active vision problem. It is however neither the only method that can be used to solve this problem, nor is the active vision problem the only problem that can be solved with visual attention. While the first point is obvious – there are dozens of methods that tackle the active vision problem which are not related to visual attention – the second point is less clear. What can be done with attention except directing the camera? Well, human selective attention is well known to be separated into covert and overt attention. Overt attention corresponds to controlling eye movements and is therefore directly related to the active vision problem. Covert attention stands for processing parts of the sensory input without looking at them with the fovea. While covert attention usually precedes eye movements, this is not always the case. For example, Johansson et al. [2001] show that simple manipulation tasks can be done without overt attention. Equally, it makes perfect sense for a robot to process some parts of the sensory input directly without steering the camera explicitly into this direction or zooming in. Here, one can also take advantage of the fact that robot sensory input is different from the human one: while the human eye produces data that has high resolution in the center and low resolution in the periphery, most cameras can capture high resolution images in the entire field of view. These images are often artificially sub-sampled to reduce the amount of data that needs to be processed. This makes it possible to perform object recognition and many other tasks directly on the input data, without controlling the camera. Active vision can be left to occasions in which this data is not sufficient, e.g. if new viewpoints of an object have to be gathered.

Let us now discuss how far we are from attentive robots such as Dobby and which parts are still missing. In the field of attention systems themselves there are still several open issues. Among these are questions like "which are the optimal features for a robot?", "how are these features integrated best?", and "how do bottom-up and top-down cues interact?". While the bottom-up part is already quite well investigated and many good solutions exist, less is known about top-down attention and existing approaches are limited to some aspects. Up to now, the prior knowledge that has been used as top-down information has mainly concentrated on two aspects of this area. First, people have used object information to search for simple objects, e.g., highlighting red regions to find fire extinguishers, [Frintrop, 2006, Navalpakkam and Itti, 2006]. Second, context information about the scene has been investigated to guide the gaze, e.g., people are likely to be on the street level of an image rather than on the sky area [Torralba et al., 2006]. However, many other cues and memories influence human perception and should also be used for attentive robots. Thus, more sophisticated knowledge about objects, people, and the situation, knowledge about typical locations of objects, as well as action cues from human interactors will strongly support the selection of regions of most potential interest.

Additionally, while this review and most existing research focuses on visual attention, other sensors can be an important source of useful information that should be exploited. Some work in this direction is our previous work on saliency detection in laser data [Frintrop et al., 2005] and the combination of visual and acoustic saliency cues for the humanoid robot iCub [Ruesch et al., 2008]. It is also important to consider that robots and humans differ considerably and that concepts that are optimized for the human brain are not necessarily optimal for a machine. While most current approaches directly transfer concepts, an important

direction of future research is to investigate how systems have to be adapted to best fit the robots' embodiment and environment.

Finally, it should be mentioned that current systems use attention mechanisms for clearly specified tasks such as landmark detection or object manipulation. While good results have been obtained in these areas, it is still a long way to obtain an attentive robot such as Dobby. Among the parts that are still missing is certainly a close interaction between different modules. In computer vision, recent work has shown that tasks such as object detection, segmentation, tracking, and categorization profit strongly from each other if the modules collaborate and share information [Leibe et al., 2008, Ess et al., 2010]. Similarly, future attentive robots will strongly profit from interacting modules. Context information and prior knowledge from other modules can enable an attentive robot to obtain better, more useful regions of interest. On the other hand, the computation of attention regions will also improve the performance of other modules since more processing resources can be provided to essential parts of the sensory input.

References

1. Definition of attention. http://en.wikipedia.org/wiki/Attention, June 2011.

Y. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. *International Journal of Computer Vision (IJCV)*, 1(4):333–356, 1988.

R. Bajcsy. Active perception vs. passive perception. In *Proc. IEEE Workshop on Computer Vision: Representation and Control*, Bellaire MI, 1985.

A. Belardinelli. *Salience features selection: Deriving a model from human evidence*. PhD thesis, Sapienza Universita di Roma, Rome, Italy, 2008.

M. Björkman and D. Kragic. Active 3D scene segmentation and detection of unknown objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, USA, 2010.

M. Bollmann, R. Hoischen, M. Jesikiewicz, C. Justkowski, and B. Mertsching. Playing domino: A case study for an active vision system. In H.I. Christensen, editor, *Computer Vision Systems*, pages 392–411. Springer, 1999.

C. Breazeal. A context-dependent attention system for a social robot. In *Proc. of the Int'l Joint Conference on Artifical Intelligence (IJCAI 99)*, pages 1146–1151, Stockholm, Sweden, 1999.

C. Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots*. PhD thesis, Department of Electrical Engineering and Computer Science. MIT, 2000.

N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 2009.

J. J. Clark and N. J. Ferrier. Modal control of an attentive vision system. In *Proc. of the 2nd International Conference on Computer Vision*, Tampa, Florida, US, Dec 1988.

A. Ess, K. Schindler, B. Leibe, and L. Van Gool. Object detection and tracking for autonomous navigation in dynamic environments. *International Journal of Robotics Research*, 29(14):1707–1725, 2010.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/, 2010.

T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3-4):143–166, 2003.

P.-E. Forssén, D. Meger, K. Lai, S. Helmer, J. J. Little, and D. G. Lowe. Informed visual search: Combining attention and object recognition. In *International Conference on Robotics and Automation*, 2008.

S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, volume 3899 of *Lecture Notes in Artificial Intelligence (LNAI)*. Springer, Berlin/Heidelberg, 2006.

S. Frintrop. The high repeatability of salient regions. In *Proc. of ECCV workshop "Vision in Action: Efficient Strategies for Cognitive Agents in Complex Environments"*, 2008.

S. Frintrop. Computational visual attention. In A. A. Salah and T. Gevers, editors, *Computer Analysis of Human Behavior (to appear)*, Advances in Pattern Recognition. Springer, 2011.

S. Frintrop and P. Jensfelt. Attentional landmarks and active gaze control for visual SLAM. *IEEE Trans. on Robotics, Special Issue on Visual SLAM*, 24(5), Oct 2008.

S. Frintrop, E. Rome, A. Nüchter, and H. Surmann. A bimodal laser-based attention system. *J. of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance in Computer Vision*, 100(1-2):124–151, Oct-Nov 2005.

S. Frintrop, M. Klodt, and E. Rome. A real-time visual attention system using integral images. In *Proc. of the 5th Int'l Conf. on Computer Vision Systems (ICVS)*, Bielefeld, Germany, March 2007.

S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. on Applied Perception*, 7(1), 2010.

D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Trans. on PAMI*, 31(6), 2009.

S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. Bradski, P. Baumstarck, S. Chung, and A. Y. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video. In *Proc. of the 20th Int. Joint Conference on Artifical intelligence (IJCAI)*, 2007.

G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Machine Vision and Applications*, 16(1):64–73, 2004.

X. Hou and L. Zhang. Saliency detection: a spectral residual approach. In *Proc. of CVPR*, 2007.

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

W. James. *The Principles of Psychology*. Dover Publications, New York, 1890.

R. Johansson, G. Westling, A. Backstrom, and J. Flanagan. Eye-hand coordination in object manipulation. *The Journal of Neuroscience*, 21(17):6917–6932, 2001.

M. Johnson-Roberson, J. Bohg, M. Björkman, and D. Kragic. Attention based active 3D point cloud segmentation. In *Proc. of the 2010 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, October 2010.

B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. of Computer Vision, Special Issue on Learning for Recognition and Recognition for Learning*, 77(1-3):259–289, 2008.

T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

D. Loach, A. Frischen, N. Bruce, and J. K. Tsotsos. An attentional mechanism for selecting appropriate actions afforded by graspable objects. *Psychological Science*, 19(12), 2008.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. of Computer Vision (IJCV)*, 60(2):91–110, 2004.

D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, D. G. Lowe, and B. Dow. Curious george: An attentive semantic robot. *Journal Robotics and Autonomous Systems*, 56(6), 2008.

A. D. Mehta, I. Ulbert, and C. E. Schroeder. Intermodal selective attention in monkeys. I: Distribution and timing of effects across visual areas. *Cerebral Cortex*, 10(4), 2000.

F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In *Proc. SPIE 46 Annual Int'l Symposium on Optical Science and Technology*, volume 4479, pages 12–23, Nov 2001.

S. Mitri, S. Frintrop, K. Pervölz, H. Surmann, and A. Nüchter. Robust object detection at regions of interest with an application in ball recognition. In *IEEE Proc. of the Int'l Conf. on Robotics and Automation (ICRA '05)*, 2005.

C. Muhl, Y. Nagai, and G. Sagerer. On constructing a communicative space in HRI. In *Proc. of the 30th German Conference on Artificial Intelligence (KI 2007)*. Springer, 2007.

Y. Nagai. From bottom-up visual attention to robot action learning. In *IEEE 8th Int'l Conf. on Development and Learning*, 2009.

C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000.

V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

S. B. Nickerson, P. Jasiobedzki, D. Wilkes, M. Jenkin, E. Milios, J. K. Tsotsos, A. Jepson, and O. N. Bains. The ARK project: Autonomous mobile robots for known industrial environments. *Robotics and Autonomous Systems*, 25(1-2): 83–104, 1998.

S. E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, Cambridge, MA, 1999.

H. Pashler. *The Psychology of Attention*. MIT Press, Cambridge, MA, 1997.

B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic. An active vision system for detecting, fixating and manipulating objects in real world. *International Journal of Robotics Research*, 29(2-3), 2010.

A. Rotenstein, A. Andreopoulos, E. Fazl, D. Jacob, M. Robinson, K. Shubina, Y. Zhu, and J.K. Tsotsos. Towards the dream of intelligent, visually-guided wheelchairs. In *Proc. 2nd Int'l Conf. on Technology and Aging*, 2007.

J. Ruesch, M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, and R. Pfeifer. Multimodal saliency-based bottom-up attention: A framework for the humanoid robot icub. In *Proc. of Int'l Conf. on Robotics and Automation (ICRA)*, 2008.

B. Schauerte, J. Richarz, and G. A. Fink. Saliency-based identification and recognition of pointed-at objects. In *Proc. of Int. Conf. on Intelligent Robots and Systems (IROS)*, 2010.

C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Transaction on Robotics*, 25(4):861–873, July 2009.

A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4), 2006.

A. M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

J. K. Tsotsos. *A Computational Perspective on Visual Attention*. The MIT Press, 2011.

J. K. Tsotsos, G. Verghese, S. Stevenson, M. Black, D. Metaxas, S. Culhane, S. Dickinson, M. Jenkin, A. Jepson, E. Milios, F. Nuflo, Y. Ye, and R. Mann. PLAYBOT: A visually-guided robot to assist physically disabled children in play. *Image and Vision Computing 16, Special Issue on Vision for the Disabled*, pages 275–292, April 1998.

S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal. Overt visual attention for a humanoid robot. In *Proc. International Conference on Intelligence in Robotics and Autonomous Systems (IROS 2001)*, pages 2332–2337, Hawaii, 2001.

J. Vogel and N. de Freitas. Target-directed attention: Sequential decision-making for gaze planning. In *Proc. of ICRA*, 2008.

D. Walther. *Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics*. PhD thesis, California Institute of Technology, Pasadena, CA, 2006.

D. Walther and C. Koch. Attention in hierarchical models of object recognition. *Computational Neuroscience: Theoretical insights into brain funciton, Progress in Brain research*, 165:57–78, 2007.

J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1(2):202–238, 1994.

T. Xu, T. Pototschnig, K. Kühnlenz, and M. Buss. A high-speed multi-GPU implementation of bottom-up attention using CUDA. In *Proc. of the International Conference on Robotics and Automation, (ICRA)*, 2009.

T. Xu, T. Zhang, K. Kühnlenz, and M. Buss. Attentional object detection of an active multi-vocal vision system. *Int. J. of Humanoid Robotics*, 7(2), 2010.