FP6-004381-MACS

MACS

Multi-sensory Autonomous Cognitive Systems Interacting with Dynamic Environments for Perceiving and Using Affordances

Instrument: Specifically Targeted Research Project (STReP)

Thematic Priority: 2.3.2.4 Cognitive Systems

D3.1.3 Saliency Detection with Visual Attention

Due date of deliverable: November 30, 2005
Actual submission date: January 6, 2005

Start date of project: September 1, 2004 Duration: 36 months

Joanneum Research (JR_DIB)

Revision: Version 1

Project co-funded by the European Commission within the Sixth Framework Programme (2002–2006)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

EU Project MACS

Deliverable 3.1.3

Saliency Detection with Visual Attention

Simone Frintrop, Martin Hülse, Erich Rome, Lucas Paletta

Number: **MACS/3/1.3**
WP: *3.1*
Status: *Version 1*
Created at: *June 30, 2005*
Revised at: *January 6, 2006*

FhG/AIS	Fraunhofer Institut für Autonome Intelligente Systeme, Sankt Augustin, D
JR_DIB	Joanneum Research Graz, A
LiU-IDA	Linköpings Universitet, Linköping, S
METU-KOVAN	Middle East Technical University, Ankara, T
OFAI	Österreichische Studiengesellschaft für Kybernetik, Vienna, A

This research was partly funded by the European Commission's 6th Framework Programme IST Project MACS under contract/grant number FP6-004381. The Commission's support is gratefully acknowledged.

Author addresses:

Dr.-Ing. Erich Rome
Fraunhofer Institut für
Autonome Intelligente Systeme
Schloß Birlinghoven
D-53754 Sankt Augustin, Germany

Fraunhofer Institut für
Autonome Intelligente Systeme
Schloss Birlinghoven
D-53754 Sankt Augustin
Germany

Tel.: +49 (0) 2241 14-2683
(Co-ordinator)

Contact:
Dr.-Ing. Erich Rome

Joanneum Research
Institute of Digital Image Processing
Computational Perception (CAPE)
Steyrergasse 9
A-8010 Graz
Austria

Tel.: +43 (0) 316 876-1769

Contact:
Dr. Lucas Paletta

Linköpings Universitet
Dept. of Computer and Info. Science
Linköping 581 83
Sweden

Tel.: +46 13 24 26 28

Contact:
Prof. Dr. Patrick Doherty

Middle East Technical University
Dept. of Computer Engineering
Inonu Bulvari
TR-06531 Ankara
Turkey

Tel.: +90 312 210 5539

Contact:
Prof. Dr. Erol Şahin

Österreichische Studiengesellschaft
für Kybernetik (ÖSGK)
Freyung 6
A-1010 Vienna
Austria

Tel.: +43 1 5336112 0

Contact:
Prof. Dr. Georg Dorffner

Contents

1 Introduction

2 Efficient Visual-Sensor processing by Attention Systems

When dealing with complex sensor data as received by cameras and 3D laser scanners the processing amount is computationally high. This is one of the biggest problems in dealing with such sensors. To enable efficient processing it is helpful to concentrate on regions of interest in the sensor data. One approach for the detection of such regions is the modeling of human visual attention. In human vision, attention helps to identify regions of relevant data and scan these regions sequentially with saccades.

Computational attention systems use psychological and neurobiological findings for the computation of regions of interest (see section 2.1). One approach of such systems is to compute different features like intensity, color and orientations in parallel by linear filters, detecting conspicuous points in each feature channel. These conspicuities are collected in a single saliency map, topographically coding the saliency of the environment. Sequentially moving the focus of attention to the most salient parts in this map enables to select regions of potential interest and to focus processing to these parts.

Currently, existing attention systems are restricted to the processing of camera data whereas data from other sensors is not considered. This is in contrast to human perception, where information of different senses is fused and a combined focus of attention is generated. As a new approach, in this project we want to integrate data from multiple sensors - camera, 3D laser scanner, manipulation sensor - to enable the utilization of their respective advantages.

Another lack of existing systems is that they usually work only bottom-up. That means, the approach is purely data-driven and only objects with strong features immediatly popping out from their environment can be detected. Top-down attention, i.e., influencing the saliency of special regions by task dependend hints, is still rarely considered in psychological and neurobiological research as well as in computational systems. Recently, some groups started research concerning this topic (see section 2.1) but many questions are still open. The further examination of top-down attention and the integration into the system will be an important part in this project.

2.1 State of the Art

Concerning visual attention, most research has so far been done in the field of *bottom-up* processing (in psychology [TG80; Wol94], neuro-biology [CS02; Pal99] and computer vision [KU85; IKN98; BMB01; SF03]). Bottom-up attention is merely data-driven and finds regions that attract the attention automatically, e.g., a black sheep in a white flock. Koch & Ullman [KU85] described the first explicit computational architecture for bottom-up visual attention. It is strongly influenced by Treisman's *feature-integration theory* [TG80] and already contains the main properties of many current visual attention systems, e.g., the one by Itti et al. [IKN98] or [BMB01; SF03]. These systems use classical linear filter operations for feature extraction, rendering them especially useful for real-world scenes. Another approach is provided by models consisting of a pyramidal neural processing architecture, e.g., the *selective tuning model* by Tsotsos et al. [TCW$^+$95]. We presented in [FNS04; FRNS05; FNSH04; MFP$^+$05; Fri05] the bottom-up part of our attention system VOCUS which is based on a standard architecture [IKN98] but differs in several aspects yielding a

considerably improved performance (see section 2.2).

While much less analyzed, there is strong neurobiological and psychophysical evidence for top-down influences modifying early visual processing in the brain due to preknowledge, motivations, and goals [Yar69; CS02; WHK+04]. However, only a few computational attention models integrate top-down information. The earliest approach is the *guided search* model by Wolfe [Wol94], a result of his psychological investigations of human visual search. Tsotsos' system considers feature channels separately and uses inhibition for regions of a specified location or those that do not fit the target features [TCW+95]. Schill et al. [SUB+01] use top-down information from a knowledge-base to select actual fixations from the fixation candidates determined by a bottom-up system. In their system, the computation of the bottom-up features is not influenced by the top-down information. Hamker performs visual search on selected images but without considering the target background [Ham04]. The closest related work is presented by Navalpakkam et al. [NRI05]; however, the region to learn is not determined automatically and exciting and inhibiting cues as well as bottom-up and top-down cues are not separated. Furthermore, quality and robustness of the system are not shown. To our knowledge, there exists no complete, well investigated system of top-down visual attention comparable to our approach.

Applications of computational attention systems are usually found in the fields of computer vision and robotics. In computer vision, most work has been done in the field of object recognition: A (usually bottom-up) attention system detects regions of interest and a classifier recognizes the content of the region [MPI01; WRKP04; SAA02; Oue03]. As a new approach, we presented in [MFP+05; Fri05] the combination of a top-down modulated attention system with classification. A different view on attention for object recognition is to determine discriminative regions of interest in objects [FSP04; PE99]. Other application scenarios in computer vision can be found in the fields of image compression [OBH+01], image matching [FB03]., image segmentation [OAHE02; OH03], object tracking [OH04] or active vision [MBHS99; Bol99; VCSS01; CF89; DPC98].

In the field of robotics, applications of computational attention systems are found for example in human-robot interaction [Bre99; HRB+04; Rae00], in object manipulation [TVS+98; BHJ+99; Rae00], in navigation [BP97; SE97] and in localization [NJW+98; OH04].

The use of attention in the combination with affordances has to our knowledge not yet been done.

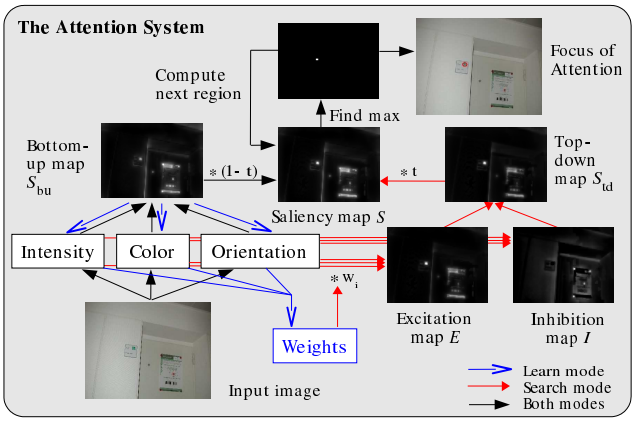
2.2 The Visual Attention System VOCUS

The computational attention system that shall be used in MACS is the system VOCUS (Visual Object detection with a CompUtational attention System). It was already introduced in [Fri05; FRNS05; MFP+05; FBR05a; FBR05b; FNSH04]. A complete overview is best obtained from [Fri05].

VOCUS consists of a bottom-up part computing data-driven saliency and a top-down part enabling goal-directed search. Global saliency is determined from both cues. An overview of the system is given in Fig. 1.

2.2.1 Bottom-up saliency

VOCUS' bottom-up part detects salient image regions by using image contrasts and uniqueness of a feature, e.g., a red ball on green grass. It was inspired by Itti et al.



Feature	weights
intensity on/off	0.001
intensity off/on	9.616
orientation 0°	4.839
orientation 45°	9.226
orientation 90°	2.986
orientation 135°	8.374
color green	76.572
color blue	4.709
color red	0.009
color yellow	0.040
conspicuity I	6.038
conspicuity O	5.350
conspicuity C	12.312

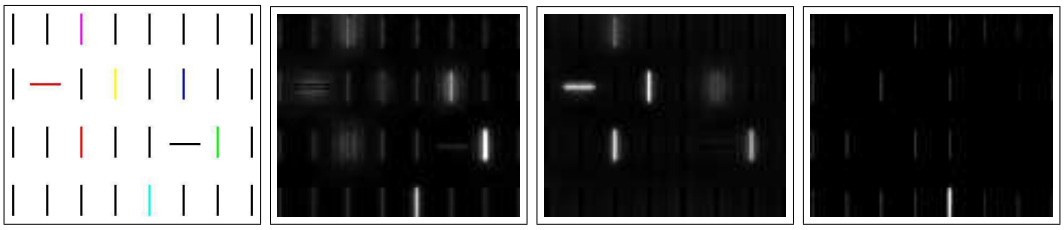
Figure 1: The goal-directed visual attention system with a bottom-up part (left) and a top-down part (right). In learning mode, target weights are learned (blue arrows). These are used in search mode (red arrows). Right: weights for target name plate.

[IKN98] but differs in several aspects resulting in considerably improved performance (see [Fri05]). The feature computations are performed on 3 different scales using image pyramids. The feature intensity is computed by *center-surround mechanisms* extracting intensity differences between image regions and their surroundings, similar to the on-center and off-center ganglion cells in the human visual system [Pal99]. In contrast to [IKN98], we compute on-off and off-on contrasts separately [Fri05; FBR05b; FRNS05]; after summing up the scales, this yields 2 intensity maps. Similar, 4 orientation maps $(0°, 45°, 90°, 135°)$ are computed by Gabor filters and 4 color maps (green, blue, red, yellow) by first converting the RGB image into the Lab color space, second determining the distance of the pixel color to the prototype color (the red map shows high activations for red regions and small ones for green regions) and third, applying center-surround mechanisms. Each feature map X is weighted with the uniqueness weight $\mathcal{W}(X) = X/\sqrt{m}$, where m is the number of local maxima that exceed a threshold t. This weighting is essential since it emphasizes important maps with few peaks, enabling the detection of *pop-outs* (outliers). After weighting, the maps are summed up to the bottom-up saliency map S_{bu}.

2.2.2 Top-down saliency

To perform visual search, VOCUS first computes target-specific weights (learning mode) and, second, uses these weights to adjust the saliency computations according to the target (search mode). We call this target-specific saliency *top-down saliency*.

2.2.2.1 Learning mode. In learning mode, VOCUS is provided with a training image and coordinates of a *region of interest (ROI)* that includes the target. The region might be the output of a classifier specifying the target or determined manually by the user. Then, the system computes the bottom-up saliency map and the *most salient region (MSR)* inside the ROI. So, VOCUS is able to decide autonomously what is important in a ROI, concentrating on parts that are most salient and disregarding the background or less salient parts. Note that this makes VOCUS also robust to small changes of the ROI coordinates.



(a) Test image (b) Excitation map (c) Inhibition map (d) Top-down map

Figure 2: A typical schematic test display (a) as being used in perceptual psychology was put into VOCUS. Some maps (b–d) of the search for the cyan vertical bar (5th in last row). The bar region is highlighted in the excitation map but the green bar shows even more activation. Only the inhibition of the green bar enables the highest activation of cyan in the top-down map.

Next, weights are determined for the feature and conspicuity maps, indicating how important a feature is for the target. The weight w_i for map X_i is the ratio of the mean saliency in the target region $m_{(MSR)}$ and in the background $m_{(image-MSR)}$: $w_i = m_{(MSR)}/m_{(image-MSR)}$ where $i \in \{1, ..., 13\}$. This computation does not only consider which features are the strongest in the target region, it also regards which features separate the region best from the rest of the image.

The learning of weights from one single training image yields good results if the target object occurs in all test images in a similar way, i.e., on a similar background and in a similar orientation. These conditions occur if the objects are fixed elements of the environment, e.g. fire extinguishers. Nevertheless, for movable objects it is necessary to learn from several training images which features are stable and which are not. This is done by determining the average weights from n training images using the geometric mean of the weights, i.e., $w_{i,(1..n)} = \sqrt[n]{\prod_{j=1}^{n} w_{i,j}}$. Instead of using all images from the training set, we choose the most suitable ones: first, the weights from one training image are applied to the training set, next, the image with the worst detection results is taken and the average weights from both images are computed. This procedure is repeated iteratively as long as the performance increases (details in [Fri05; FBR05b]).

2.2.2.2 Search mode. In search mode, we determine a top-down saliency map that is integrated with the bottom-up map to yield global saliency. The top-down map itself is composed of an excitation and an inhibition map. The excitation map E is the weighted sum of all feature and conspicuity maps X_i that are important for the learned region, i.e., $w_i > 1$. The inhibition map I shows the features more present in the background than in the target region, i.e., $w_i < 1$:

$$\begin{array}{rcll} E & = & \sum_i (w_i * X_i) & \forall i : w_i > 1 \\ I & = & \sum_i ((1/w_i) * X_i) & \forall i : w_i < 1 \end{array} \quad (1)$$

The top-down saliency map S_{td} results from the difference of E and I and a clipping of negative values: $S_{td} = E - I$. To make S_{td} comparable to S_{bu}, it is normalized to the same range. I, E, and S_{td} are depicted in Fig. 2, showing that the excitation map as well as the inhibition map have an important influence.

The global saliency map S is the weighted sum of S_{bu} and S_{td}. The contribution of each map is adjusted by the top-down factor $t \in [0..1]$: $S = (1-t) * S_{bu} + t * S_{td}$. For $t = 1$, VOCUS considers only target-relevant features (pure top-down). For a lower t, salient bottom-up cues may divert the focus of attention, an important mechanism in human attention: a person suddenly entering a room immediately catches our attention. Also colored cues divert the search for non-colored objects as shown in [The04]. Determining appropriate values for t depends on the system state, the environment and the current task; this is beyond the scope of this article and will be tackled when integrating our attention system into robotic applications.

After the computation of the global saliency map S, the most salient region is determined by *region growing* starting with the maximum of S. Finally, the focus of attention (FOA) is directed to this region. To compute the next FOA, this region is inhibited and the selection process is repeated.

3 Attention guided Exploration and Search of Affordances

In order to not drown in perceptions and affordances, a mechanisms is needed to select the relevant parts from the input data.

VOCUS provides the selection mechanism to extract relevant regions. As we mentioned it before, such an extraction process can be based on a bottom-up or a top-down saliency.

With respect to an affordance guided robot-environment interaction we utilize the extraction mechanisms of VOCUS to let the robot run in two modes: exploration and search mode.

3.1 Exploration mode

In the exploration mode, VOCUS provides regions that might be of potential interest and that shall be investigated further by the robot, no matter which saliency (top-down or bottom-up) is used.

Due to the outcome of a specific interaction with an object in the selected region of interest, the features of the region can be linked to specific affordances. Basically, VOCUS delivers regions of interests. Additional features or feature qualities of the selected region can be derived by several other methods too, e.g. SIFT. Hence, an affordance of an object can be described by different feature qualities.

On the other hand, a feature quality can describe different affordances. An affordance is determined by the outcome of a specific robot-object interaction. If another interaction is applied then another affordances may be evaluated.

In such a way, the selected regions of interests deliver many concrete examples of robot-object interactions. These examples can be represented as concrete instances of an affordance and feature relation.

A generalization of these concrete relations can be generated by decision trees, as they are introduced in Deliverable 3.1.2. One decision tree represents a hypothesis about a relation between a specific affordance and feature qualities.

One may view a knowledge base of a robot as a set of different decision tress. Each affordance is represented by one tree. The set of trees is incrementally built up through the VOCUS guided robot-environment interaction.

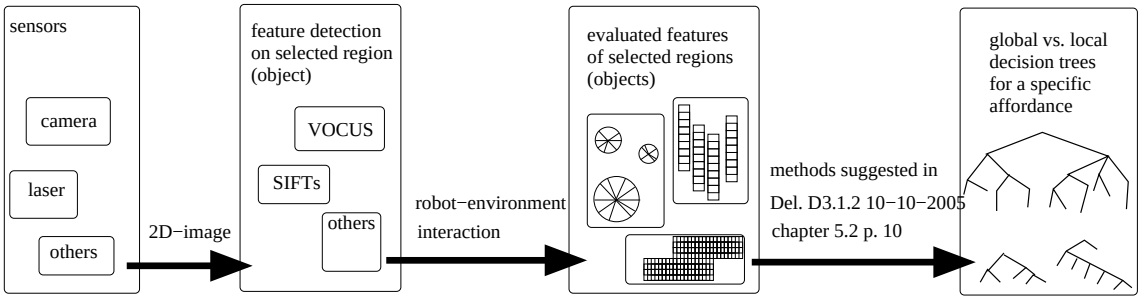


Figure 3: Schema of the generation of a set of decision trees utilized to represent hypotheses about specific affordance-feature relations. These relations can base on different sensor qualities (camera, laser) and different feature classes derived by different methods (e.g. VOCUS, SIFT).

A decision tree can base on features derived by all available vision methods (VOCUS, SIFTS, etc.) or can base exclusively on features provided by a single method. The latter type we call a local and the former a global decision tree.

Furthermore, one may consider feature detection based on other sensor qualities than the camera images which VOCUS is using for its selection process. In our case, SIFT and VOCUS features can be derived from each sensor mode, and all these sensory data can be represented as 2-dimensional images. Hence, Laser data (representing proximity, brightness as grey value images) can be used by VOCUS in the same way as camera images.

To sum up, in the exploration mode the knowledge base of the robot can be built-up with multi-sensor data as well as multi-modal because decision trees can be generated from different feature qualities as well as from different sensor qualities (Fig. 3).

3.2 Search mode

Once a set of decision trees is established, they can be used to extract specific features of objects and regions which may provide the corresponding affordance.

Assume a task the robot has to solve. For this task objects offering a specific affordance are needed. The corresponding decision tree contains the values of the VOCUS features which are associated with the needed affordance. Hence, if the top-down saliency of VOCUS is initialized with these values then VOCUS delivers image regions containing those features. Hence, the selected object / region offers the required affordance.

The selected region could be further evaluated with respect to other features or sensor qualities. That possibility of multi-modal and multi-sensor evaluation of an object for a given affordance increases the robustness of the robot interaction, since wrong assumptions about the object-affordance relation become less probable.

But there may be situations where a derived object-affordance relation is wrong. In such a case the supposed outcome of the robot-object interaction fails at a certain point. Such a failure can be used in two ways. It may initiate an adaptation of the knowledge base, i.e. the corresponding decision tree. On the other hand, it can simply initiate a new selection of other VOCUS feature values from the corresponding decision tree. These values lead to new regions of interests containing objects, which may have the needed affordance.

3.3 Learning attention patterns

In an affordance analysis setup that would be even more complex than the one described above, one could consider to focus attention not only on a single region but even on any *set* of regions in the field of view. These regions can make up a specific pattern of geometrically related local features that would themselves provide the opportunity to cue to any considerable affordance relation determined by the physical interaction of the robot with the environment. A methodology to tackle the discrimination of attention patterns by means of sequential analysis of attended regions has been described in Deliverable D5.2.1, Section 3.2. Within that Section, the learning of relevant attention patterns is outlined in a mathematical framework of Markov decision processes (MDP, POMDP).

It is planned to integrate the sequential attention framework into the VOCUS system in order to take advantage of attending to more complex affordance cues within future work.

4 Summary and Outlook

In this document we have specified how an autonomous robot can build-up a knowledge base of object-affordance relations due to its interaction within its environment. The knowledge base is represented by a set of decision trees. One decision tree can contain multi-sensor data and features derived by several feature detection methods. The "knowledge acquisition" is driven by the attention system VOCUS as well as VOCUS is utilized to select appropriate objects / regions once a specific affordance is needed.

To demonstrate the attention driven knowledge acquisition we integrate the camera, laser scanner and other features detection methods in a general software framework.

Furthermore, we test the performance of affordance cueing (Deliverable 3.1.2 Affordance recognition from visual cues) extended by the VOCUS features.

As a first proof of concept, we initialize a robot with a knowledge base and pre-selected affordances and let that system run in the search mode. Depending on which affordance is selected the robot should approach different objects / regions.

Finally, we apply the exploration mode to a simple scenario, started with bottom-up saliency and evaluate the resulting set of decision trees.

References

[BHJ+99] 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.

[BMB01] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active-vision system. *IEEE Trans. on PAMI*, 23(12):1415–1429, 2001.

[Bol99] M. Bollmann. *Entwicklung einer Aufmerksamkeitssteuerung für ein aktives Sehsystem*. PhD thesis, Universität Hamburg, Germany, 1999.

[BP97] Shumeet Baluja and Dean Pomerleau. Expectation-based selective attention for visual monitoring and control of a robot vehicle. *Robotics and Autonomous Systems*, 22(3-4):329–344, December 1997.

[Bre99] Cynthia Breazeal. A context-dependent attention system for a social robot. In *Proc. of the Sixteenth International Joint Conference on Artifical Intelligence (IJCAI 99)*, pages 1146–1151, Stockholm, Sweden, 1999.

[CF89] J. J. Clark and N. J. Ferrier. Control of visual attention in mobile robots. In *IEEE Conference on Robotics and Automation*, pages 826–831, 1989.

[CS02] Maurizio Corbetta and Gordon L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews*, 3(3):201–215, 2002.

[DPC98] J. A. Driscoll, R. A. Peters, and K. R. Cave. A visual attention network for a humanoid robot. In *Proc. of the International Conference on Intelligent Robots and Systems (IROS '98)*, pages 1968–1974, 1998.

[FB03] Friedrich Fraundorfer and Horst Bischof. Utilizing saliency operators for image matching. In *Proc. of the International Workshop on Attention and Performance in Computer Vision (WAPCV '03)*, pages 17–24, Graz, Austria, April 3 2003.

[FBR05a] S. Frintrop, G. Backer, and E. Rome. Goal-directed search with a top-down modulated computational attention system. In *Proc. of the Annual meeting of the German Association for Pattern Recognition (Jahrestagung der Deutschen Arbeitsgemeinschaft für Mustererkennung) DAGM 2005 (accepted)*, Lecture Notes in Computer Science (LNCS), Conference: Wien, Austria, Sept. 2005. Springer.

[FBR05b] S. Frintrop, G. Backer, and E. Rome. Selecting what is important: Training visual attention. In *Proc. of the 28th German Conference on Artificial Intelligence (KI 2005) (accepted)*, Lecture Notes in Computer Science (LNCS), Conference: Koblenz, Germany, Sept. 2005. Springer.

[FNS04] Simone Frintrop, Andreas Nüchter, and Hartmut Surmann. Visual attention for object recognition in spatial 3D data. In *International Workshop on Attention and Performance in Computer Vision (WAPCV '04)*, pages 75–82, Conference: Prag, Czech Republic, Mai 2004.

[FNSH04] Simone Frintrop, Andreas Nüchter, Hartmut Surmann, and Joachim Hertzberg. Saliency-based object recognition in 3D data. In *Proc. of the International Conference on Intelligent Robots and Systems (IROS '04)*, pages 2167 – 2172, Conference: Sendai, Japan, September 2004.

[Fri05] Simone Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search.* PhD thesis, University of Bonn, Germany, to appear 2005.

[FRNS05] Simone Frintrop, Erich Rome, Andreas Nüchter, and Hartmut Surmann. A bimodal laser-based attention system. *accepted for Journal of Computer Vision and Image Understanding (CVIU), Special Issue on Attention and Performance*, 2005.

[FSP04] Gerald Fritz, Christin Seifert, and Lucas Paletta. Attentive object detection using an information theoretic saliency measure. In Lucas Paletta, John K. Tsotsos, Erich Rome, and Glyn W. Humphreys, editors, *Proc. of the 2nd international workshop on attention and performance in computational vision (WAPCV '04)*, pages 136–143, Conference: Prague, Czech Republic, May 2004.

[Ham04] F. Hamker. Modeling attention: From computational neuroscience to computer vision. In *Proc. of WAPCV'04*, pages 59–66, Prague, Czech Republic, May 2004.

[HRB+04] 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.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20(11):1254–1259, 1998.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.

[MBHS99] B. Mertsching, M. Bollmann, R. Hoischen, and S. Schmalz. The neural active vision system NAVIS. In B. Jähne, H. Haussecke, and P. Geissler, editors, *Handbook of Computer Vision and Applications*, volume 3, pages 543–568. Academic Press, 1999.

[MFP+05] Sara Mitri, Simone Frintrop, Kai Pervölz, Hartmut Surmann, and Andreas Nüchter. Robust object detection at regions of interest with an application in ball recognition. In *IEEE 2005 Proc. of the International Conference on Robotics and Automation (ICRA '05)*, pages 126–131, Conference: Barcelona, Spain, April 2005.

[MPI01] F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In *Proc. SPIE 46 Annual International Symposium on Optical Science and Technology*, volume 4479, pages 12–23, Nov 2001.

[NJW+98] 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.

[NRI05] Vidhya Navalpakkam, Jim Rebesco, and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, 2005.

[OAHE02] N. Ouerhani, N. Archip, H. Hügli, and P. J. Erard. A color image segmentation method based on seeded region growing and visual attention. *International Journal of Image Processing and Communication*, 8(1):3–11, 2002.

[OBH+01] N. Ouerhani, J. Bracamonte, H. Hügli, M. Ansorge, and F. Pellandini. Adaptive color image compression based on visual attention. In *Proc. of the International Conference of Image Analysis and Processing (ICIAP 01)*, pages 416–421. IEEE Computer Society Press, 2001.

[OH03] Nabil Ouerhani and Heinz Hügli. MAPS: multiscale attention-based presegmentation of color images. In *4th International Conference on Scale-Space Theories in Computer Vision*, volume 2695, pages 537–549. Springer Verlag, Lecture Notes in Computer Science (LNCS), 2003.

[OH04] Nabil Ouerhani and Heinz Hügli. AttentiRobot: a visual attention-based landmark selection approach for mobile robot navigation. In Lucas Paletta, John K. Tsotsos, Erich Rome, and Glyn W. Humphreys, editors, *Proc. of the 2nd international workshop on attention and performance in computational vision (WAPCV '04)*, pages 83–89, Conference: Prague, Czech Republic, May 2004.

[Oue03] Nabil Ouerhani. *Visual Attention: From Bio-Inspired Modeling to Real-Time Implementation*. PhD thesis, Institut de Microtechnique Université de Neuchâtel, Switzerland, 2003.

[Pal99] Stephen E. Palmer. *Vision Science, Photons to Phenomenology*. The MIT Press, Cambridge, MA, 1999.

[PE99] L. Pessoa and S. Exel. Attentional strategies for object recognition. In J. Mira and J.V. Sachez-Andres, editors, *Proc. of the International Work-Conference on Artificial and Natural Neural Networks (IWANN '99)*, volume 1606 of *Lecture Notes in Computer Science (LNCS)*, pages 850–859, Alicante, Spain, 1999. Springer.

[Rae00] Robert Rae. *Gestikbasierte Mensch-Maschine-Kommunikation auf der Grundlage visueller Aufmerksamkeit und Adaptivität*. PhD thesis, Technische Fakultät der Universität Bielefeld, Germany, 2000.

[SAA02] A. Salah, E. Alpaydin, and L. Akrun. A selective attention based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(3):420–425, 2002.

[SE97] Christian Scheier and Steffen Egner. Visual attention in a mobile robot. In *Proc. of the IEEE International Symposium on Industrial Electronics*, pages 48–53, 1997.

[SF03] Yaoru Sun and Robert Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.

[SUB+01] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C Zetzsche. Scene analysis with saccadic eye movements: Top-down and bottom-up modeling. *Journal of Electronic Imaging*, 10(1):152–160, 2001.

[TCW+95] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *AI*, 78(1-2):507–545, 1995.

[TG80] Anne M. Treisman and Garry Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[The04] Jan Theeuwes. Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, 11:65–70, 2004.

[TVS+98] 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.

[VCSS01] Sethu Vijayakumar, Jörg Conradt, Tomohiro Shibata, and Stefan 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.

[WHK+04] J. M. Wolfe, T.S. Horowitz, N. Kenner, M. Hyle, and N. Vasan. How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research*, 44:1411–1426, 2004.

[Wol94] J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1(2):202–238, 1994.

[WRKP04] Dirk Walther, Ueli Rutishauser, Christof Koch, and Pietro Perona. On the usefulness of attention for object recognition. In Lucas Paletta, John K. Tsotsos, Erich Rome, and Glyn W. Humphreys, editors, *Proc. of the 2nd international workshop on attention and performance in computational vision (WAPCV '04)*, pages 96–103, Conference: Prague, Czech Republic, May 2004.

[Yar69] A. L. Yarbus. *Eye Movements and Vision.* Plenum Press (New York), 1969.