Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

A Biologically-Inspired Vision Architecture for Resource-Constrained Intelligent Vehicles year = 2010

Thomas Michalke, Jannik Fritsch, Christian Goerick

2010

Preprint:

This is an accepted article published in Computer Vision and Image Understanding, Special Issue: Intelligent Vision systems for Computer Vision and Image Understanding. The final authenticated version is available online at: https://doi.org/[DOI not available]

A Biologically-Inspired Vision Architecture for Resource-Constrained Intelligent Vehicles

Thomas Michalke [a,*], Jannik Fritsch [b], Christian Goerick [b]

[a] *University of Technology Darmstadt, Institute for Automatic Control D-64283 Darmstadt, Germany*

[b] *Honda Research Institute Europe GmbH, D-63073 Offenbach, Germany*

Abstract

The use of computer vision for assisting the driver dates back to first research projects in the 80's, but only recently the progress in vision research and the increase in computational power have resulted in actual products. Although impressive from the robustness point of view, these systems are optimized for specific problems and at best perform reactive tasks like, e.g., lane keeping assistance. However, for a better understanding of generic traffic situations and for assisting the driver in the full range of his actions, integrated and more flexible approaches are needed. In this contribution we propose a vision system that in important aspects is inspired by the human visual system for organizing the different visual routines that need to be carried out. The presented system searches for biological motivation in case classical engineering-based approaches cannot do better or fail. Using a tunable visual attention system and state-of-the-art perception algorithms, the system is capable of analyzing the scenery for task-relevant information in order to provide the driver with assistance in dangerous situations. Our main research focus is on the design of general mechanisms (i.e., not domain or task-specific) that lead to a certain observable behavior without being explicitly designed for this behavior. Using this principle, we aim at developing easily extensible driver assistance systems. The system components are evaluated on a complex inner-city scene and on further real world data. We demonstrate the performance of the integrated vision system in a construction site setup. A traffic jam within the construction site results in a dangerous situation that the system has to identify in order to warn the driver. Different from other systems the detection of the dangerous situation is based on the vision channel alone. Radar is only used to assign distance data to visually detected objects. The contribution represents an important intermediate stage for future, more cognitive driver assistance systems.

Key words: Vision Architecture, Driver Assistance, Top-Down Visual Attention

* Corresponding author. Email address: thomas_paul.michalke@daimler.com

1 INTRODUCTION

Among the various possible applications of vision systems, the task of driver assistance is highly interesting as it implicitly contains the challenge of understanding a dynamic scene and is at the same time of great commercial and social importance. The goal of building such an intelligent vision system can be approached from two directions: either searching for the best engineering solution or taking the human as a role model. In the latter case, research results from disciplines like, e.g., psychophysics or neurobiology can be used to guide the vision system design. While it may be argued that the quality of an engineered system in terms of isolated aspects like, e.g., object detection or tracking, is often sound, the solutions lack the necessary flexibility. Small changes in the task and/or environment often lead to the necessity of redesigning the whole system. Considering the human vision system, nature has managed to realize a highly flexible system capable of adapting to severe changes in the task and/or the environment. Hence one of our main design goals is to implement a system able to accomplish new tasks without adding modules or changing the system's structure. Equally, we aim at getting inspiration from the underlying principles of the human vision system and not directly at engineering efforts to attain its measurable abilities.

It is important to note that we do not focus on building a close 'psychophysical model' of the human vision system that models all its known aspects as close as possible. Among other things, said models are useful for predicting or explaining measurements in psychophysical studies with humans. Different from such a global paradigm, we mimic functionality-related findings of the human visual pathway in cases known classical approaches do not perform better, are restricted in their flexibility, or perform less robust. Put differently, the contribution aims at realizing a 'computational model' of the human vision system that allows robust, real-time operation in a real-world environment (please refer to [1] for a comprehensive discrimination between computational and psychophysical models of the human vision system). The envisioned system modulates and parameterizes submodules without being explicitly designed for specific tasks of a scenario.

Aiming at going beyond standard industrial computer vision applications, there is an increasing emphasis in the computer vision community on building so-called cognitive vision systems [2] (i.e., systems that work according to or get inspiration from human information processing principles) suitable for solving complex vision tasks. One important principle in cognitive systems is the existence of top-down links in the system, i.e., informational links from stages of higher to lower knowledge integration. Top-down links are believed to be a prerequisite for fast-adapting biological systems living in changing environments (see, e.g., [3]).

Returning to the car domain, constraints like, e.g., lane markings and traffic rules restrict the environmental complexity and ease the driving task considerably. Still, vision-based driver assistance functionalities developed up to now are mainly capable of dealing with simple traffic situations. While this already resulted in specialized commercial products improving driving safety (e.g., the 'Honda Intelligent Driver Support System' [4] which helps the driver to stay in the lane and maintain the right distance to the preceding car), the problem of developing a generic vision system for advanced driver assistance, i.e., capable of operating in all kinds of challenging situations, is still unsolved.

One possible way to achieve this goal is to realize a task-dependent perception using top-down links. In this paradigm, the same scene can be decomposed in different ways depending on the current task. A promising approach for decomposing the scene is to use a high-performance attention system that can be modulated in a task-oriented way, i.e., based on the current context. For example, while driving at high speed, the central field of the visual scene becomes more important than the surrounding.

Aiming towards such a task-based vision system, this contribution describes a first instance of a vision architecture that is being developed as perceptual front-end of an Advanced Driver Assistance System (ADAS). The proposed system provides a framework that enables the task-dependent tuning of visual processes via object-specific weighting of input features of the attention system. The system generates an appropriate system reaction in dangerous situations (autonomous braking). In major parts, its architecture is inspired by findings in the human visual system and organizes the different functionalities in a similar way. For a first proof of concept, we focus on assisting the driver during a critical situation in a construction site. For the analysis of the attention system, we evaluated the construction site scenario as well as a challenging inner-city traffic scene to illustrate the performance gain of the top-down approach in a more complex environment. Furthermore, additional images of real world traffic scenes are used to evaluate different system modules. The overall system achieves real-time performance on a prototype car and is evaluated offline on 10 construction site streams and online on 60 documented test drives. The obtained results demonstrate the feasibility and benefits of top-down attention and the chosen architectural approach in a complex ADAS.

The contribution is organized as follows: In Section 2 we relate our work to research on visual attention systems and existing car vision architectures. Subsequently, Section 3 provides an overview of the system architecture and goes into the details of the visual attention processes. Evaluation results for the most crucial system modules as well as the overall system performance measured in an experimental setup are given in Section 4. The contribution ends with a summary and an outlook on future work in Section 5.

2 RELATED WORK

In the following section, an introduction and overview of existing computational attention models is given. Since the focus of this contribution is on system-related aspects that allows for building the vision part of an Advanced Driver Assistance System running in real-world scenarios, the remaining Section will then focus on existing vision systems suitable for the vehicle domain.

Facilities for controlling and managing traffic are always visually conspicuous. For example, lane markings are white on a typically dark road and traffic signs or traffic lights have bright colors. According to that, in many countries flashy advertisement is prohibited in the proximity of roads. The said examples exploit a key aspect of the human visual processing - the principle of early selection. With vision being the most important sensory modality of humans having the highest information density, the named principle significantly accelerates the processing of vision data. More specifically, the abundance of visual stimuli in the world is prefiltered or preselected early to match the restricted cognitive capacity of the human brain. In plain words, the principle of early selection suppresses sensor data that is not relevant to the current needs or goals of the system causing a colorful, bright traffic sign to visually pop-out in a traffic scenario (see [5] for details). For realizing said early selection principle the human disposes of the so-called attention mechanism, which preselects certain scene elements.

More specifically, the human vision system filters the high abundance of environmental information by attending to scene elements that either pop out most in the scene (i.e., objects that are visually conspicuous) or match the current task best (i.e., objects that are compliant to the current internal state or need/task of the system), while suppressing the rest. For both attention guiding principles psychophysical and neurological evidence exists (see [6,7]).

Furthermore, the psychophysical experiments of Simons and Chabris [8] impressively showed that the task has a modulating effect on attention. The gathered results were formalized in the concept of 'inattentional blindness'. In their experiments, participants did not notice unexpected events (like a black gorilla walking through an indoor scene) when the task (counting ball contacts of a white basketball team) involved features complementary to the unexpected events (see Fig. 1).

Following this principle, technical vision systems have been developed that prefilter a scene by decomposing it into its features (see [9]) and recombining these to a saliency map that contains high activation at regions that differ strongly from the surroundings (i.e., *bottom-up (BU) attention*, see [10] for the underlying psychophysical attention model). A well-known computational

Fig. 1. Psychophysical study conducted by [8] marking the human visual attention as strong mediator between the world and our perception of the world.

BU attention model for saliency calculation is the approach by Itti et al. [11] that is used in a number of implemented systems. More recent system implementations additionally include the modulatory influence of task relevance into the saliency (i.e., *top-down (TD) attention*), see [12] as one of the first and [13–20] as more recent and probably most influential approaches. Typically, the named systems apply dynamic weights to different processing stages, with the task to find a specific object within a predominantly static indoor scene. A more complete view on a possible architecture for a vision system that incorporates task-dependent visual attention is given by Navalpakkam and Itti [14,21]. The proposed architecture combines top-down (TD) and bottom-up (BU) influences by using TD weights on the calculated BU features. However, there is no separation between the untuned BU-saliency map and the calculated TD saliency maps allowing a weighted combination, which would ensure the preservation of BU influence in all system states. The system is evaluated mainly on static indoor scenes and a few static outdoor scenes. There are only few attention-based vision systems that use a motion feature (see [22–24]). Given the importance of motion in the human visual perception, we see integrating the influence of scene dynamics on attention as a key issue to realize robust human-like vision systems.

In these systems, instead of scanning the whole scene in search of certain objects in a brute force way, the use of TD attention allows a full scene decomposition despite restraints in computational resources. In principle, the vision input data is serialized with respect to the importance to the current task. Based on this, computationally demanding processing stages higher in the architecture work on prefiltered data of higher relevance, which saves computation time and allows complex real-time vision applications.

Endowing a vision architecture for an intelligent car with similar, task-based attention can result in a gain of performance with minimal additional resource requirements (see Section 4).

In [24], we chose the two related top-down attention systems of Navalpakam [14] and Frintrop [1] for a detailed structural and functional comparison, since these impacted our work most. In [24], we also described and tested approaches

that make our attention system particularly appropriate for the real-world vehicle domain.

However, numerous other psychophysical and computational attention models exist (please refer to [1,?,?,?] for a comprehensive overview of the latest developments in attention research and [25] for an overview of related psychophysical studies).

Regarding typical real-world scenarios in the vehicle domain, the robustness of biological attention systems is difficult to achieve, given e.g., the high variability of scene content, changes in illumination, and scene dynamics. Most computational attention models do not show real-time capability and are mainly tested in a controlled indoor environment on artificial scenes. Important aspects discriminating real-world scenes from indoor and artificial scenes are the dynamics in the environment (e.g., changing lighting and weather conditions, dynamic scene content) as well as the high scene complexity (e.g., cluttered scenes). Dealing with such scenarios requires a strong system adaptation capability with respect to changes in the environment.

During the vision system design we aimed at a computational efficient system implementation for online use on vehicles. The overall system should be flexible, meaning that a new system task should not lead to the necessity of realizing new modules or a structural redesign of the whole system. Getting our inspiration from biology we therefore aimed at a system that exhibits specific properties without being specifically designed for these properties (e.g., our system is able to locate the horizon edge or detect fast moving objects or red traffic signs without being explicitly designed for these tasks). More specifically, the design goals of our TD attention sub-system comprised the development of an object- and task-specific tunable saliency map suitable for the real-world scenarios in the car domain.

In this contribution, we focus on important conceptual issues crucial for closing the gap between artificial and natural attention systems operating on real-world scenes. We show the feasibility of our approach on vision data from the car domain. The described TD tunable attention system is used as front-end of the vision system of an advanced driver assistance system (ADAS), whose architecture is in important parts inspired by the human brain.

Coming to attention-related research in the vehicle domain, the task-dependent nature of gazing has also been proven while steering a car. Recently, it was shown in [26] that the performance for dangerous situation detection (a colored motorcycle veering into the vehicle's path) strongly depends on the feature match between the current distracting visual task and the unexpected obstacle. In another example, the gaze of drivers in a virtual environment was examined [27]. The results show that the performance in detecting stop

signs is heavily modulated by context (i.e., top-down) factors and not only by bottom-up visual saliency.

Turning to the domain of complete vision systems developed for ADAS, there have been few attempts to incorporate aspects of the human visual system. With respect to attention processing, a saliency-based traffic sign detection and recognition system was demonstrated in [28]. In terms of complete vision systems, one of the most prominent examples is a system developed in the group of E. Dickmanns [29]. It uses several active cameras mimicking the active nature of gaze control in the human visual system. However, the processing framework is not closely related to the human visual system. Without a tunable attention system and with TD aspects that are limited to a number of object-specific approaches for classification, no dynamic preselection of image regions is performed. A more biologically inspired approach has been presented by Färber [30]. However, their publication as well as the recently started German Transregional Collaborative Research Centre 'Cognitive Automobiles' [31] address mainly human inspired behavior planning whereas our work described here focuses more on the task-dependent perception aspects.

The only other known vision system approach that attempts to explicitly model aspects of the human visual system is described by [32]. The system is somewhat related to the here presented ADAS. However, published after our work (see, e.g., [33]), the approach allows for a simple attention-based decomposition of road scenes but without incorporating object knowledge or pre-knowledge. Additionally, the overall system organization is not biologically inspired and hence shows limitations in its flexibility.

In contrast to the here presented ADAS, a tendency of most large-scale research projects like, e.g., the European PreVENT project [34] is the decomposition of the overall functionality into many building blocks and combining these blocks into subsets for solving isolated tasks. While this 'divide and conquer' approach does lead to impressive results in specific settings, we believe the challenge of integrating all these functionalities into a coherently working flexible system is not yet solved.

In the following Section, the implemented driver system is described in detail. It contains the following community-related novelties:

- A driver assistance system on a prototype vehicle was implemented that allows autonomous emergency braking on highways based on vision as the major cue,
- The realized driver assistance system is based on a computational attention system as generic front-end of all visual processing allowing task-dependent scene decomposition and interpretation in real-time.

On a functional level the following novelties could be reached:

- Computationally efficient decomposition of the Gabor filter response (a specific saliency feature described in Section 3) in on-off and off-on components, allowing a gain in selectivity for the attention system,
- A subfeature normalization procedure that assures the comparability of BU and TD attention without loosing information about the absolute signal amplitude,
- A biologically motivated homeostasis approach (see Section 3) for making diverse modalities comparable.

3 SYSTEM ARCHITECTURE

In the following, after defining our design goals, an overview of the implemented vision system structure for driver assistance is given. Subsequently, crucial system parts are described in more detail.

3.1 Design Goals

The following list gives an overview of the design goals that drove our system development. We aimed at:

- Realizing a generic system structure whose modules and links between modules can be modulated (i.e. parameterized) online,
- A system that realizes a specific task-dependent processing without being explicitly designed for these tasks,
- A system that explicitly takes the human as a role model on the micro level (mimicking human signal processing principles, e.g., specific filter kernels that were measured in the brain, retina-like color processing) and on the macro level (mimicking the organization and combination of signal flows in the brain),
- Searching for biological inspiration is not a global paradigm in our design process, in the sense that we mimic functionality-related findings of the human visual pathway only in cases known classical approaches do not perform better.

In order to reach these goals the following cognitive principle were applied or gave us motivation:

- Top-down links (i.e., links from higher levels of system integration that modulate lower levels of system integration) that allow a task-dependent modulation of lower signal processing principles,

- The principle of Inhibition of Return, since it increases the efficiency of visual search,
- A visual attention system, since it allows for generic, task-dependent scene decomposition and top-down tuning,
- The principle of early selection that improves the relevancy of input data of higher system levels based on a task-dependent preselection of input data,
- A separation into 'what' and 'where' processing pathways similar to the assumed organization of the human brain.

3.2 Overview

The overall architecture concept to realize task-based visual processing is depicted in Fig. 2. It contains a distinction between a 'what' and a 'where' processing path, somewhat similar to the known properties of the human visual system where the 'dorsal' and 'ventral' pathways are typically associated with these two functions. Among other things, the 'where' pathway in the human brain is believed to perform the localization and coarse tracking of a small number of objects that are relevant for the current task. This tracking is performed by the human visual system without focusing the eye gaze on individual objects to be tracked [35], i.e., tracking does not require high resolution. In contrast, the 'what' pathway considers the detailed analysis of a single spot in the image. In the human visual system this is intimately bound to the current eye gaze, as the human eye possesses a high resolution in the central 2-3° (foveal retina area) of the visual field only.

In our vision system, the eye gaze is performed virtually as the camera mounted in the car has a constant resolution in the complete field of view. Changing the eye gaze is therefore equivalent to shifting the processing to another spot of the input image. This spot is analyzed by the classifier (higher part of the 'what' pathway) in full resolution while the whole image is analyzed in the attention sub-system (lower part of the 'what' pathway) as well as in the 'where' path in lower resolution. Processing in these two pathways is believed to occur in parallel in the human brain, but their intertwinings are as yet not known in too much detail. We here adopt the idea of continuously tracking a small number of objects in each image of the incoming visual stream to coarsely represent the current scene and at the same time acquiring more detailed information on one additional object. We therefore have two analysis processes running in parallel in our system (see Fig. 2).

The detailed organization of the two processing streams in our architecture concept is as follows: The input image is analyzed in the 'what' path (depicted on the left in Fig. 2) for salient locations using a variety of visual features including orientation, intensity, color, and motion. This visual atten-

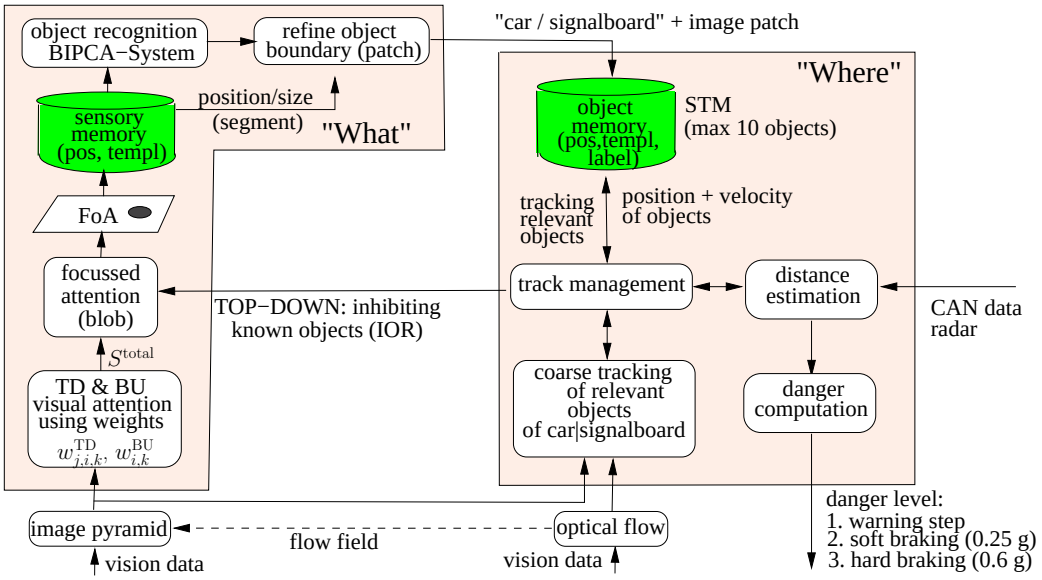


Fig. 2. Architecture concept of our vision-based driver assistance system.

tion combines bottom-up (BU) and top-down (TD) pathways and is described in more detail in Section 3.3. The resulting saliency map S^{total} is modulated by suppressing image regions that contain known objects, i.e., that have been detected earlier. The system stores all detected objects in a so-called Short Term Memory (STM) that provides the position information of known objects as top-down link. The suppression of saliency areas is also known as Inhibition of Return (IoR) in the human visual system [36] and is included in numerous other computational attention models to facilitate the visual attention-based search. However, the manner how the IoR is used in computational attention models differs in some aspects from findings in the human vision system (see [36]). Among other things, the IoR: 1) Was found to be attached to environmental locations, instead of retinal coordinates as in our attention system, 2) Also depends on planed fixation points and scene locations, which suggests numerous additional modulating influences. Still, major IoR-related properties found in the human vision system are in accordance with the here presented system. For instance, as stated in [36] the IoR: 1) Can last for several seconds, 2) Was demonstrated to be attached to moving objects, which suggests that it is coupled to brain regions that support object tracking, 3) Facilitates visual foraging (i.e., visual search). The stated performance gain in visual search caused by using the IoR approach and the influence on the STM will be shown in Section 4.

A simple maximum search is used on the resulting saliency map to find the currently most salient point in the scene. The Focus of Attention (FoA) is determined by region growing on the overall saliency map using the most salient point as an anchor. For the named approach, we got inspiration by [1]. The saliency-based region growing approach is generic, meaning that it has the advantage of being independent from the type of the TD search target. De-

pendency of the segmentation algorithm from the search target would require the development of object-specific segmentation approaches and would hence restrict the approach to a limited number of objects. The drawback of said approach is that the segmentation might fail in case a salient object background or another close and salient object is present. Since the used classifier also supports roughly segmented input data that can contain other objects or background, the approach is still robust enough for using it on real world data on our prototype car (see evaluation results in Section 4). However, we plan to integrate a so-called figure-ground segmentation approach that is based on Learning Vector Quantization in order to solve the named challenges (see [37] for details). The FoA (pos, RoI) is then fed to the fast feedforward object recognition system (see Section 3.6).

The resulting object position and the image segment (pos, template) are stored together with the object label (i.e., the object class) in the STM in order to be coarsely tracked in subsequent images in the 'where' path. Before insertion, it is checked whether the new object can be associated to a known object, in this case the object already stored in the STM is updated. Concluding one iteration, for all objects in the STM a distance estimation (dist) is calculated based on fusing measurements from radar, depth from familiar object size (also called depth from object knowledge, see [5]) and from bird's eye view [38] using an Extended Kalman Filter (see Section 3.7). This information is stored in a separate egocentric representation that is directly suitable for assessing the current danger level in the scene and generating a warning if necessary.

All objects contained in the STM are constantly tracked in the 'where' path based on an appearance-based tracker that uses a second order motion model for prediction and a local correlation step for the refinement of the new object positions. In each iteration the position is updated in the STM and a new object template is stored. In case the prediction does not match (no good correlation found) the object is deleted from the STM and therefore its position will not be inhibited any longer in the 'what' pathway. Consequently, the attention will be focused on the missing object in one of the next images if the object is still present and salient. This way, all objects being recognized and behaving as predicted are coarsely tracked while the 'what' attention is always focused on new objects and objects behaving unexpectedly. Note that the rather simple tracking method is sufficient for many applications in the automotive domain where most objects are rigid (e.g., a car) and therefore the main appearance changes are caused by small translations and scaling. In a subsequent ADAS implementation, we introduced a 3D prediction approach that allows the simple compensation of the camera vehicle's ego motion (see [39]), which in turn makes the tracking more robust.

The key aspect of our architecture lies in the introduction of top-down aspects (like, e.g., task-dependent tunable attention generation via sets of weights and,

in parallel, inhibiting known object positions predicted by tracking) resulting in the ability to cope with highly dynamic traffic scenes using limited computational resources. The top-down tunable attention system is a key principle of our ADAS, since such preprocessing leads to a considerable reduction of scene complexity by restricting further processing steps to image regions that are interesting according to the current system task.

3.3 Attention Sub-System

In the following, our biologically motivated attention system is described that is one of the key aspects of the contribution at hand. The description is done in a rather compressed fashion. More details on the design goals as well as the five explicit novelties of the realized attention system can be found in [24].

A simplified sketch of the visual attention sub-system is depicted in Fig. 3. It consists of a number of features that are extracted from the image on 5 scales derived from a Gaussian image pyramid starting from 256 × 256 pixels. The Gaussian image pyramid was calculated by low pass filtering and dyadic downsampling. The lower right half of Fig. 3 shows the bottom-up processing of the different features to obtain bottom-up conspicuity maps that are combined to form the bottom-up saliency S^{BU}. The conspicuity maps represent the different modalities (e.g., color, motion, edges) supported by the system. In the upper right half the top-down processing is shown where all subfeature maps are weighted and combined into the top-down conspicuity maps. All top-down conspicuity maps are combined into the object-specific top-down saliency map S^{TD} and a nonlinear operator is applied to cut off negative values. The overall saliency map S^{total} is calculated by linearly combining the normalized top-down S^{TD} and bottom-up saliency maps S^{BU} depending on the current task of the ADAS using parameter λ (see Eq. (1)).

$$S^{\mathrm{total}} = \lambda S^{\mathrm{TD}} + (1 - \lambda) S^{\mathrm{BU}} \tag{1}$$

The resulting saliency map S^{total} is passed on to the FoA generation.

As features (modalities) we currently use *odd* and *even Gabor filters* [40] with additional on-center/off-center separation (for details see Section 3.5) in 4 orientations, *Difference of Gaussians filters* (DoG) as on-center/off-center, *motion* from differential images and the biologically motivated *RGBY color* space as color opponent and double color opponent [1]. In sum we use M=7 feature types (modalities) that in turn are composed of 136 subfeature maps. However, it is important to note that not all modalities can be applied in BU and TD pathway (e.g., plain colors contain no BU information and are hence not supported in the BU pathway, see Fig. 3). The feature map responses are

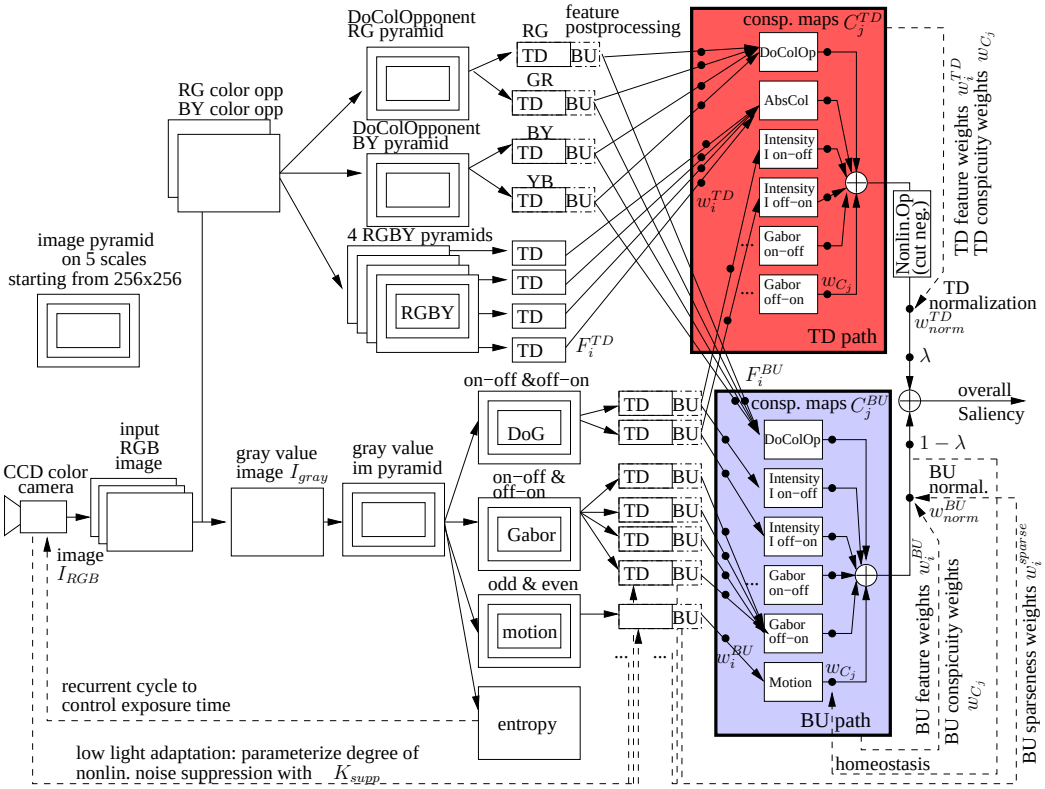


Fig. 3. Simplified sketch of the visual saliency that is part of our attention sub-system (showing for each feature only one output).

passed through a preprocessing step that consists of normalization, squaring, and nonlinear noise suppression by a sigmoidal function (see Section 3.4 for further details). In addition to combining these features to obtain a bottom-up saliency map [11,10], we also compute top-down saliency maps using object-specific feature map weights. The object-specific weights are inspired by [15,41] in the way the weights are obtained: During a supervised training stage, the feature map activations of an object (region of interest(RoI)) are compared to the feature map activations in its surrounding/background (see Fig. 4 for a visualization). From this comparison, the relative importance of a feature (its signal-to-noise (SNR) ratio) can be determined. For each trained object and feature channel F_i we therefore get a top-down weight w_i^{TD} that is proportional to how well the feature channel i is able to discriminate the object from its surrounding:

$$w_i^{\mathrm{TD}} = \begin{cases} \frac{m_{RoI,i}}{m_{rest,i}} & \forall \frac{m_{RoI,i}}{m_{rest,i}} \geq 1 \\ -\frac{m_{rest,i}}{m_{RoI,i}} & \forall \frac{m_{RoI,i}}{m_{rest,i}} < 1 \end{cases} \tag{2}$$

$$\text{with} \quad m_{\{\mathrm{RoI,rest}\},i} = \frac{\sum_{\forall u,v \in \{\mathrm{RoI,rest}\}} F_i(u,v)}{\text{size region } \{\mathrm{RoI,rest}\}}$$

$$\text{and} \quad F_i(u,v) = \begin{cases} F_i(u,v) & \forall (u,v), F_i(u,v) \geq \phi \\ 0 & \text{else} \end{cases}$$

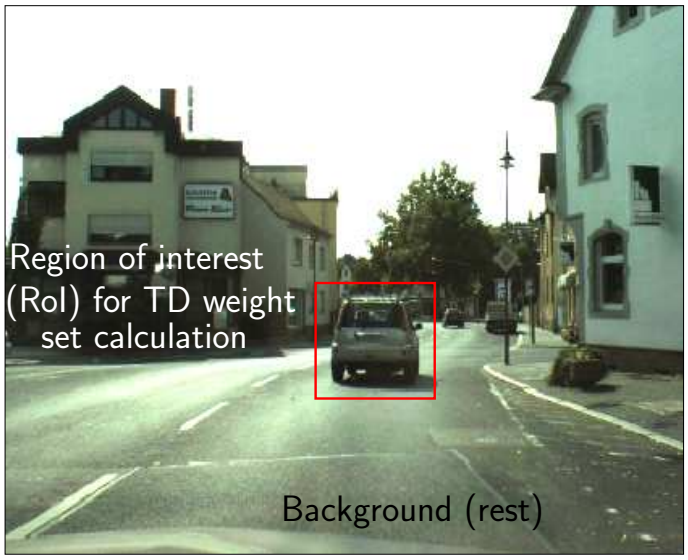


Fig. 4. Visualization of the object training region (RoI) for TD weight calculation against the background (rest).

According to Eq. (2) matching features are boosted (excitation) and irrelevant features are suppressed (inhibition). As visualized in Fig. 3, the $j = 1..M$ TD conspicuity maps result from a weighted combination of the N_j TD subfeature maps within a certain feature type j (see Eq. (3)).

$$C_j^{\mathrm{TD}} = \sum_{i=1}^{N_j} w_{i,j}^{\mathrm{TD}} F_{i,j}^{\mathrm{TD}} \tag{3}$$

It is important to note that the performance gain of this approach compared to most of the attention systems described Section 2 lies in the explicit inhibition of non-target regions combined with a high feature selectivity. The conspicuity maps C_j^{TD} are combined to an object-specific top-down saliency map S^{TD}

by modality specific weights w_{C_j} (conspicuity weights) that are proportional to the confidence one can assign to the modality j in the current scene. This is done dynamically depending on, e.g., the current weather or lighting conditions (see Section 3.4). The TD saliency results from a weighted sum of 6 different conspicuity maps (even Gabor on-off, odd Gabor off-on, DoG on-off, DoG off-on, RGBY color opponent, RGBY double color opponent), see Eq. (4).

$$S^{\mathrm{TD}} = \sum_{j=1}^{6} w_{C_j} C_j^{\mathrm{TD}} \tag{4}$$

In addition, we also calculate a biased bottom-up saliency map (see Eq. (6)) by combining all feature maps weighted with their specific bottom-up weights w_i^{BU} resulting in the weighted sum of 6 BU modalities (Gabor and DoG as for TD, RGBY double color opponent, motion), see Eq. (5):

$$C_j^{\mathrm{BU}} = \sum_{i=1}^{N_j} w_{i,j}^{\mathrm{BU}} F_{i,j}^{\mathrm{BU}} \tag{5}$$

$$S^{\mathrm{BU}} = \sum_{j=1}^{6} w_{C_j} C_j^{\mathrm{BU}} \tag{6}$$

As w_i^{BU} we choose a set of weights that shows good performance for most situations in the vehicle domain. In the object-unspecific bottom-up path no inhibition takes place, since its purpose is to evaluate the general unspecific saliency of a scene.

The individual bottom-up feature maps F_i^{BU} are additionally preprocessed by a pop-out operator that globally amplifies maps with a small number of maxima and attenuates maps with many maxima [11]. The pop-out operator multiplies the feature maps with a dynamic factor w_i^{sparse} computed at runtime (see Eq. (7)). The factor is inversely proportional to the number of pixels that are near the maximum of the feature map. Additionally, w_i^{sparse} is increased by a factor of 2 for each higher (i.e., smaller) scale level s in the image pyramid. As higher levels tend to contain more pixels fulfilling the threshold defined in the denominator of Eq. (7), the increase of the factor s maintains the comparability of scales:

$$w_i^{\mathrm{sparse}} = \sqrt{\frac{2^s}{\sum\limits_{\forall u,v \text{ with } F_i(u,v)>\xi} F_i(u,v)}} \text{ for } s = [0,4] \text{ and } \xi = 0.9 \cdot Max(F_i) \tag{7}$$

By applying this operator, the bottom-up path is designed to amplify feature

maps that show few maxima, i.e., that are sparse. In consequence, feature maps containing image regions that pop out are boosted. It is of crucial importance that the top-down feature maps do not pass a similar pop-out step, since by tuning the top-down weights, we aim at finding objects based on feature conjunctions. The individual feature map responses for the searched objects might only reach medium values, whereas the combination of all relevant maps leads to a strong response in the resulting saliency map. This explicit differentiation is not made in other top-down attention systems, which leads to a performance loss, as was shown in [24].

For weighting the feature maps we currently use TD weight sets for signal boards and cars ($w^{\mathrm{TD}}_{i,sigboard}$ and $w^{\mathrm{TD}}_{i,car}$) that were calculated in a supervised training step. In a more recent version of our ADAS these weights are computed dynamically at runtime (see [39]). In our current (see [42]) and future work it is envisioned to use attention weights to track and even learn new objects.

3.4 Feature Postprocessing, Normalization, and Homeostasis of Conspicuity Maps

In the following the feature postprocessing is described, focusing on the used normalization procedure (see also Fig. 6) that makes our approach different from other attention systems. All subfeatures are normalized to the theoretical maximum value that can be expected for the specific subfeature map (not the current maximum on the map). For example, for DoG and Gabor this is done by determining the filter response for the ideal input pattern, maximizing the filter response. Figure 5 shows the ideal DoG and 0° even Gabor input pattern (suitable for the filter kernels depicted in Fig. 7a and Fig. 7c). This

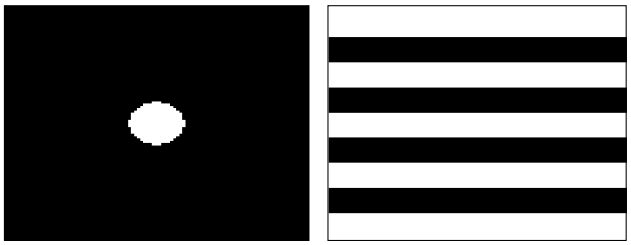


Fig. 5. Input patterns that maximize the filter response. The maximum of this filter response is used for normalization: (a) Ideal DoG input pattern, (b) Ideal 0° even Gabor input pattern.

procedure ensures comparability between subfeatures of one modality while preserving information about the absolute feature amplitude. Now, the signal power is calculated by squaring, after which a sigmoid function is applied for noise suppression. A parameter K_{supp} shifts the sigmoid function horizontally,

which influences the degree of noise suppression and the sparseness of the resulting subfeature maps.

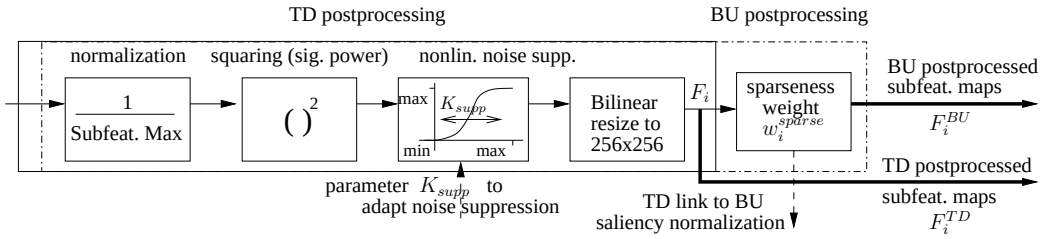


Fig. 6. Postprocessing of feature maps

The subfeature normalization procedure ensures intra-feature comparability, but for the overall combination, comparability between modalities (i.e., conspicuity maps) is required as well. We solve the normalization problem of the conspicuity maps by dynamically adapting the conspicuity weights w_{C_j} for weighting the BU and TD conspicuity maps C_j^{BU} and C_j^{TD}. This concept mimics the homeostasis process (see e.g., [43]), which we understand as the property of a biological system to regulate its internal processes in order to broaden the range of environmental conditions in which the system is able to survive. More specifically, the $\tilde{w}_{C_j}(t)$ are set to equalize the activation on all $j = 1..M$ BU conspicuity maps (see Equation (8)), taking only the N_j pixel over the threshold $\xi = 0.9 \cdot Max(C_j^{\mathrm{BU}})$ into account. Exponential smoothing (see Equation (9)) is used to fuse old conspicuity weights $w_{C_j}(t-1)$ with the new optimized ones $\tilde{w}_{C_j}(t)$. The parameter α sets the velocity of the adaptation and could be adapted online dependent on the gist (i.e., basic environmental situation) via a TD link. In case of fast changes in the environment α could be set high for a brief interval, e.g., while passing a tunnel or low in case the car stops. Additionally, we use thresholds for all M conspicuity maps based on a sigma interval of recorded scene statistics to avoid complete adaptation to extreme environmental situations.

$$\tilde{w}_{C_j}(t) = \frac{1}{\frac{1}{N_j} \sum\limits_{\forall u,v \text{ with } C_j^{\mathrm{BU}}(u,v) > \xi} C_j^{\mathrm{BU}}(u,v)} \quad \text{and} \quad \xi = 0.9 \cdot Max(C_j^{\mathrm{BU}}) \tag{8}$$

$$w_{C_j}(t) = \alpha \tilde{w}_{C_j}(t) + (1-\alpha) w_{C_j}(t-1) \quad \text{for} \quad j = 1..M \tag{9}$$

Before combining the BU and TD saliency maps using the parameter λ (see Eq. (1)) a final normalization step takes place. Like the subfeature and conspicuity maps, the saliency maps are normalized to the maximum expected value. For this we have to step back through the attention sub-system taking into account all weights (w_i^{sparse}, w_i^{BU}, w_i^{TD}) and the internal disjointness/conjointness of the features to determine the highest value ($v_{\mathrm{max},j}^{\mathrm{BU}}$ and $v_{\mathrm{max},j}^{\mathrm{TD}}$) a single pixel can achieve in each BU and TD conspicuity map j. We

define a feature as internally disjoint (conjoint), when the input image is decomposed without (with) redundancy in the subfeature space. In other words the recombination of disjoint (conjoint) subfeature maps of adjacent scales or orientations is equal to (bigger than) the decomposed input image. Since DoG and Gabor are designed to be internally disjoint between scales and orientations (see Chapter 2) the maximum pixel value on a conspicuity map j is equal to the maximum of the product of all subfeature and/or sparseness weights of the subfeatures it is composed of (w_i^{sparse} and w_i^{BU} for BU as well as w_i^{TD} for TD). Motion is conjoint between scales, therefore we sum up the products of all subfeature motion weights w_i^{BU} and their corresponding w_i^{sparse} to get the maximally expected value on the motion conspicuity map. The contribution of the color feature to the saliency normalization weight is similar but more complex.

Since appart from DoG and Gabor there is disjointness between conspicuity maps the maximum possible pixel values for all BU and TD conspicuity maps, calculated as described above, are multiplied with the corresponding w_{C_j} and added to achieve the normalization weights w_{norm}^{TD} and w_{norm}^{BU} for the TD and BU attention (see Eq. (10) and Fig. 3 for the position the normalization weights are applied). Using this approach, w_{norm}^{TD} will adapt when the TD weight set changes (see Eq. (11)).

$$w_{norm}^{\text{BU}} = \frac{1}{\sum_{j=1}^{M} k_j w_{C_j} v_{\max,j}^{\text{BU}}} \tag{10}$$

$$w_{norm}^{\text{TD}} = \frac{1}{\sum_{j=1}^{M} k_j w_{C_j} v_{\max,j}^{\text{TD}}} \tag{11}$$

With:

$$k_j = \begin{cases} 0.5 \text{ for } & j \in \{\text{DoG}, \text{Gabor}\} \\ 1 \quad \text{for } & j \notin \{\text{DoG}, \text{Gabor}\} \end{cases}$$

It is important to note that DoG and Gabor features are conjoint, meaning that they represent the same signal characteristics. Put differently the conspicuity maps for DoG and Gabor are not independent. As discussed in Chapter 2 using both DoG and Gabor is still helpful, since the signal decomposition is different for both filter types. The conjointness is taken into account in the attention normalization procedure in Eq. (10) and (11) in the form of the factor k_j that decreases the integral influence of DoG and Gabor on the overall attention.

Using this approach w_{norm}^{TD} adapts when the TD weight set w_i^{TD} changes, yielding a TD saliency map S^{TD} that is comparable to S^{BU} for all object-specific TD weight sets.

In order to yield high hit rates in TD search, the features of an attention system need high selectivity to provide as much supporting and inhibiting maps as possible. At the same time, high efficiency is needed due to constraints in computational resources. An approach fulfilling these demands is the separation of the DoG filter in on-center (called on-off in the following) and off-center selectivity (off-on) as is emphasized in [1] (see Fig. 7a and Fig. 7b). To realize such an on-off/off-on separation the DoG filter response is separated into its positive and negative part, which is equivalent to the computationally more demanding usage of the two different filter kernels depicted in Fig. 7a and Fig. 7b. Coming to the Gabor filter, dividing the complex Gabor filter response into its real and imaginary part allows the efficient separation into edge and line selective responses (equivalent to separately filtering with an odd and even Gabor kernel, depicted in Fig. 7c-f). In addition to this well-known concept, we transfer the DoG on/off-center concept to the Gabor filter and separate the odd and even Gabor responses into their positive and negative parts. For example, an on-off versus off-on even Gabor separation allows for the efficient separation of white street markings from shadows on the street and an on-off/off-on separation for odd Gabor allows for the crisp suppression of the sky edge present in most scenes in the car domain.

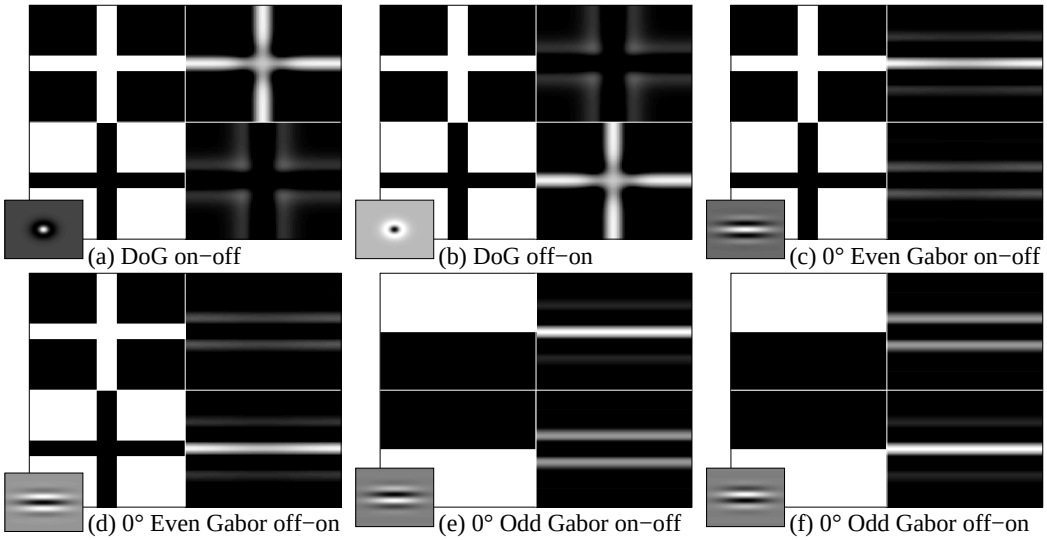


Fig. 7. Application of filter kernels on simple test images (negative filter response is cut off). Both the 2 DoG features (a),(b) and the 4 Gabor features (c)-(f) are realized with 1 filter operation each. Every image shows on the left the input test images and on the right the respective filter response for the filter kernel in the bottom left corner.

In sum, 4 different Gabor-based features are derived from one filtering step. Each of these 4 Gabor features consists of 20 independently weighable sub-

feature maps (4 orientations on 5 scales each). The well-known concept of decomposition into scales via the usage of image pyramids allows efficient image filtering. In contrast to other attention approaches, we weight the subfeatures on all scales independently (e.g., no weighting on scale level [15]). In addition, the usage of the motion feature calculated from difference images allows the system to detect and separate between slow and fast moving objects.

3.6 Object Recognition

For object classification, we use a appearance-based approach, where we perform the classification only on the image segment provided by the FoA segmentation. Note that for object recognition the original image resolution of 800×600 pixels is used, i.e., the object position and size provided by the saliency system are transformed appropriately.

The object recognition module is based on a biologically motivated processing architecture proposed in [44]. It uses a strategy similar to the hierarchical processing in the ventral pathway of the human visual system by creating a classification hierarchy. Unsupervised learning is used for the lower levels of the hierarchy to determine general features that are suitable for representing arbitrary objects robustly with regard to local invariance transformations like local shift and small rotations. Only at the highest level of the hierarchy object-specific learning is carried out, i.e., only this layer has to be trained for different objects. This architecture can be applied to the difficult case of segmentation-free recognition that we have to deal with, as the saliency segmentation only provides a rectangular image segment and no object-specific segmentation.

Training is done by presenting several thousand color image segments with changing backgrounds for back views of cars and signal boards (see also [45]). The learning algorithm automatically extracts the relevant object structures and neglects the clutter in the background. The output of the classifier is the identity/class of the recognized object and a confidence value where a threshold is used to reject object hypotheses with a low confidence. The threshold is chosen so that only a small number of false positives can occur for cars, as a wrong car detection could lead to a false emergency braking. If a car is not recognized due to the high threshold, it is stored in the STM as unknown and tracked rather shortly for N frames before it is removed from the STM. Subsequently, if the car is still a salient object, a new FoA will be generated and recognition is performed again. As now the car may be closer due to the ego motion of our vehicle, the image patch may be larger and therefore may have a higher confidence resulting in a correct recognition.

As described in Section 3.2, with the 'what' pathway, the presented system

uses a cascade of attention-based object detection followed by an appearance-based object classification. According to [46], object recognition in human perception is organized in a similar way. As argued above, the central hypothesis regarding the here presented attention-based preselection is that it saves computation time and lowers the number of false positive classifications due to the high relevancy of input data at the classifier stage. However, the question arises if in terms of computational demands, the approach is superior to an exhaustive classification of the whole image (e.g., by classifying overlapping image patches). As argued in [1], in case of a complex and thereby slow classifier, the advantages of an attention system are obvious. Since in the vehicle domain false detections might have severe consequences, with [44] a reliable and hence complex classifier was applied in the presented system.

Even for applications that allow the usage of fast (and less reliable classifiers), as Viola-Jones (see [47]) the usage of an attention system saves computational resources, as was shown in [1]. The results gathered by the author show that already in case of more than 1 object class, the computation time needed by the attention system is compensated by the need of fewer classifier cycles. Furthermore, based on numerous experiments, [1] could show that the number of false classifications is reduced in case an attention system for preselecting image regions is used as compared to applying exhaustive classification.

3.7 Depth Cues

The current ADAS uses four independent depth sources (see Fig. 9) that are combined using weak fusion (see [48]). Weak fusion combines the depth sources based on the reliability of the specific cues. It is realized here using an Extended Kalman Filter (EKF) that combines the depth cues at each time step via dynamic weights depending on static predefined sensor variances and the current availability of the depth cues (as not every cue is available in each time step). The EKF uses a second order process model for its prediction step that models the relevant kinematics in the car domain (velocity and acceleration). The resulting depth values are used to assign depth to detected objects in the image.

Depth from radar (Radio Detecting and Ranging) is obtained from a commercial standard vehicle equipment sensor, which delivers sparse point-wise measurements of low longitudinal but higher lateral uncertainty (for an example see Fig. 9b). Radar sensors evaluate the reflections (echoes) of bundled micro wave beams (typically between 400 MHz and 80 GHz). More specifically, the time of flight t_{tof} is used to determine the object distance Z_{radar} (see Eq. (12)). For measuring the time of flight the individual beam packages must be marked and recognized, which can be done by modulation and demodula-

tion of the signal amplitude, frequency or phase. The object velocity v_{dop} is determined based on the Doppler shift Δf (see Eq. (13)).

$$Z_{\text{radar}} = \frac{c_0 \cdot t_{\text{tof}}}{2} \tag{12}$$

With: c_0 ... velocity of propagation (speed of light) $\approx 300000 \frac{km}{s}$

t_{tof} ... time of flight (to the object and back)

$$v_{\text{dop}} = \frac{c_0 \cdot \Delta f}{2 f_0} \tag{13}$$

With: c_0 ... velocity of propagation (speed of light) $\approx 300000 \frac{km}{s}$

Δf ... measured Doppler frequency shift

f_0 ... carrier frequency

Using radar sensors, the object distance and velocity can hence be measured with independent approaches. Different from visual sensors, radar is very robust against changing weather conditions, which makes it an important cue that increases the system robustness.

Depth from bird's eye view: For computing the distance of objects that are positioned on the drivable path the *bird's eye view* is used. The bird's eye view is a metric representation of the scene as viewed from above (see Fig. 8a). The cue is able to detect and estimate the distance of objects present on the ego vehicle's and neighboring lane (as opposed to the perspective image). Working on this representation for estimating object distances has the advantage that the cumbersome non-linear projection from 3D world coordinates to the 2D image plane (see Eq. (14) and (15)) is intrinsically compensated. As such, world position coordinates can directly be assigned to a detected object without further processing. Furthermore, by this transformation, the detection of lanes and objects can be realized easier than when working on the projected camera image, since expectations regarding typical metric lane widths can be integrated easily into the algorithm. The bird's eye view is calculated on the undistorted pixels v and u based on Eq. (14) and (15) by inverse perspective mapping of the 3D world points X, Y, and Z (see Fig. 8b for a visualization of the used coordinate system) to the 2D (u,v) image plane. Equation (14) and (15) use the 3 camera angles θ_X, θ_Y, and θ_Z, the 3 translational camera offsets t_1, t_2, t_3 (see Fig. 8b), the horizontal and vertical principal point u_0 and v_0 as well as the horizontal and vertical focal lengths f_u and f_v (focal lengths that are normalized to the horizontal and vertical pixel size respectively). The equations describe how to map a 3D position of the world to the

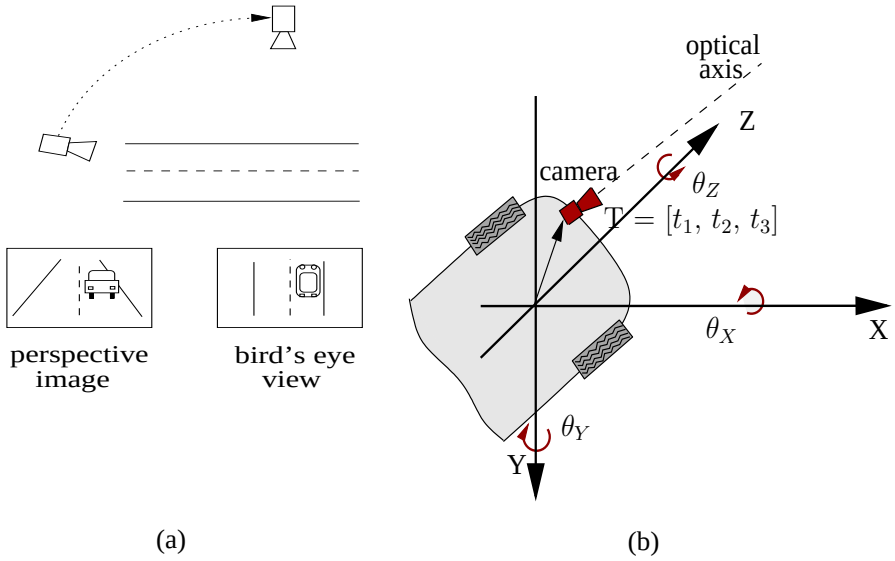


Fig. 8. (a) Visualization of the bird's eye view, (b) Coordinate system and position of the camera.

2D image plane (refer to [38]). More specifically, only the image pixels (u,v) that are needed to get the metric bird's eye view (i.e., the XZ-plane) dense are mapped, which also leads to low computational demands. The usage of inverse perspective mapping makes the inversion of Eq. (14) and (15) obsolete, when computing the bird's eye view.

$$u = -f_u \frac{r_{11}(X\text{-}t_1) + r_{12}(Y\text{-}t_2) + r_{13}(Z\text{-}t_3)}{r_{31}(X\text{-}t_1) + r_{32}(Y\text{-}t_2) + r_{33}(Z\text{-}t_3)} + c_u \tag{14}$$

$$v = -f_v \frac{r_{21}(X\text{-}t_1) + r_{22}(Y\text{-}t_2) + r_{23}(Z\text{-}t_3)}{r_{31}(X\text{-}t_1) + r_{32}(Y\text{-}t_2) + r_{33}(Z\text{-}t_3)} + c_v \tag{15}$$

$$R = R_X R_Y R_Z = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \tag{16}$$

$$\begin{aligned} r_{11}^2 + r_{12}^2 + r_{13}^2 - 1 &= 0 \\ r_{21}^2 + r_{22}^2 + r_{23}^2 - 1 &= 0 \\ r_{31}^2 + r_{32}^2 + r_{33}^2 - 1 &= 0 \\ r_{11}r_{21} + r_{12}r_{22} + r_{13}r_{23} &= 0 \\ r_{11}r_{31} + r_{12}r_{32} + r_{13}r_{33} &= 0 \\ r_{21}r_{31} + r_{22}r_{32} + r_{23}r_{33} &= 0 \end{aligned} \tag{17}$$

As can be seen in Eq. (14) and (15) the 3D world position coordinates X, Y, and Z of all image pixels (u,v) are needed. By using a monocular system, one dimension (the depth Z) is lost. A solution to this dilemma is the so-called *flat plane assumption*. Here, for all pixels in the image, the height Y is set to 0. Based on this, only objects in the image with $Y = 0$ (especially, the street we are interested in) are mapped correctly to the bird's eye view, while all the other regions are stretched to infinity in the bird's eye view (for example the car in Fig. 9d).

Now, a vertical grow algorithm with dynamic thresholds searches for discontinuities in the bird's eye view and assigns a distance value to them (see Fig. 9d).

In the rectified image (i.e., the image is remapped to be equivalent to an image with all 3 camera angles zero) the following direct relation between the vertical pixel value v and the depth Z_{birds} exists (see Eq. (18)).

$$Z_{\text{birds}} = \frac{f_v \; t_2}{(v - v_0)} \tag{18}$$

With:

t_2 ... camera height above the ground
v_0 ... the vertical principal point
v ... vertical pixel position that shows significant contrast change
f_v ... Normalized focal length

Depth from object knowledge calculates the distance of an object Z_{obj} (see Eq. (19)) using knowledge about the area the object covers on the image chip (width W_{pixel} and height H_{pixel}), the width and height of the object in the real world drawn from experience (W_{real} and H_{real}) as well as the intrinsic parameters of the sensor (α_u = focal length/pixel width and α_v = focal length/pixel height). A prerequisite for depth from object knowl-

edge is a reliable segmentation algorithm. Currently we use histogram-based segmentation on an image region that is pre-segmented by our region growing algorithm working on the saliency (see Fig. 9c)

$$Z_{obj,W} \approx \frac{W_{real}\ \alpha_u}{W_{pixel}} \quad \text{and} \quad Z_{obj,H} \approx \frac{H_{real}\ \alpha_v}{H_{pixel}} \tag{19}$$

Depth from Stereo Disparity: The perception of *stereoscopic depth* is based on the interpretation of the differences between the projected images of both eyes (so-called parallax). An isolated point in the 3D world is projected to slightly different positions on the retina of both eyes, since these have a horizontal distance, the so-called basic distance. The horizontal shift between the images is called lateral *disparity*, see [49]. In addition to the lateral disparity, other flavors of disparity exist (see [49]) that can also cause an impression of depth - still the lateral disparity seems to be the most important disparity-related depth cue and is therefore also in the focus of the following reflections. For detecting lateral disparity (for simplification called disparity in the following) the detection of correspondences between the left and right eye is necessary. Here, ambiguities are possible, due to differences in illumination and partial occlusion between both images. Especially, local regions of low texture can lead to the well-known aperture problem, which is also a challenge for the optical flow computation (refer to [50]). Furthermore, differences and changes in the internal optical parameters of both eyes exist that influence the projections and hence the detected lateral disparity. Still, the human vision system can cope with these challenges by continuous adaptation mechanisms. How these challenges are solved by the human vision system is largely unknown. Designing a technical stereo system that closely mimics the processing steps in the brain is therefore not possible up to now. However, the engineered approaches show sound results, but have limitations also.

Based on the disparity image the 3D world position X, Y, and Z for all image pixels can be computed using Eq. (20) (see Fig. 9a), (21), and (22). The equations result from a transformation of Eq. (14) and (15), setting all camera angles to zero, since the disparity computation was done on rectified images.

$$Z_{\text{stereo}}(u,v) = \frac{f_u B}{D(u,v)} + t_3 \tag{20}$$

$$Y_{\text{stereo}}(u,v) = \frac{Z(v-v_0)}{f_v} + t_2 \tag{21}$$

$$X_{\text{stereo}}(u,v) = \frac{Z(u-u_0)}{f_u} + t_1 \tag{22}$$

With: B... basic distance between the left and right camera's principal point
f_u, f_v... normalized focal length [in pixels]
$D(u,v)$... disparity
u_0, v_0... principal point
t_1, t_2, t_3... translational camera offset

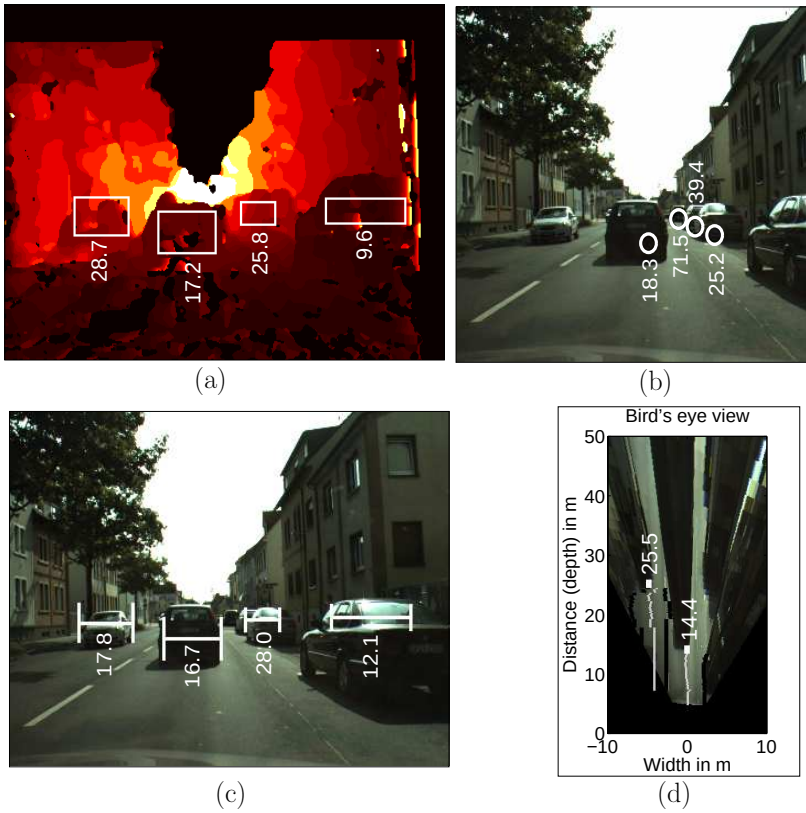


Fig. 9. Used depth cues: Depth from (a) Stereo disparity, (b) Radar, (c) Object knowledge, (d) Bird's eye view.

4.1 Evaluation of System Modules

Normalization - Comparable TD and BU saliency maps: The used feature normalization (described in Section 3.4) prevents noise on the saliency maps and ensures the preservation of the absolute level of feature activation. Using a TD weight set that supports object-specific features our normalization hence ensures that the TD map will show high activation if and only if the searched object is really present. Figure 12a shows based on a construction site scenario as depicted in Fig. 11a that the maximum attention value on the TD saliency map for cars rises when the car comes into view, which would not be present using the normalization approach found in literature (see, e.g. [11]).

The influence combining the now comparable TD and BU saliency maps for cars and reflection posts (e.g., useful for unmarked road detection as done in [51]) as trained search objects is depicted in Tab. 1, showing that TD improves the search performance considerably. It is important to note that besides an exchange of the training images no modification in the system structure is required, when changing the search object. For evaluation the measures *average FoA hit number* ($\overline{Hit}$) and *average detection rate* ($\overline{DRate}$) were calculated. While $\overline{DRate}$ is the ratio of the number of found task-relevant objects to the overall number of task-relevant objects, $\overline{Hit}$ states that the object was found on average with the $\overline{Hit}$ 'th generated FoA. Hence the smaller $\overline{Hit}$ the earlier an object is detected see [1] for a more detailed definition of these measures). The choice of training images has only small influence on the search performance as the comparable results for different sets of training images in Tab. 1 show.

The evaluation shows highest hit numbers and detection rates for pure TD search ($\lambda = 1$). However, it is important to note that pure TD search would lead to a suppression of unexpected objects (inattentional blindness, see Section 2) and would hence potentially cause dangerous situations. The default value for λ was hence set to 0.5 for the online tests. This setting allows the simultaneous detection of a specific object class (in this case vehicles) and other salient objects (as, e.g., the horizon edge, signal boards or unexpected dangerous objects in the path). In an ADAS that succeeded the here presented system, we concentrated on improving the system design in order to allow the simultaneous detection of multiple object classes based on the presented attention system (see [39]).

The presented results support the generic nature of the TD tunable attention

Target	# Test images (objects)	# Training im	$\overline{Hit}$ ($\overline{DRate}$) $\lambda = 0$ (BU)	$\lambda = 0.5$ (BU & TD)	$\lambda = 1$ (TD)
Cars	54 (58)	54 (self test)	3.06 (56.9%)	1.56 (93.1%)	1.53 (100%)
Train. set 1		3		1.87 (89.7%)	1.82 (96.6%)
Train. set 2		2		1.90 (84.5%)	1.76 (93.1%)
Train. set 3		3		1.96 (82.8%)	1.94 (93.1%)
Train. set 4		3		1.84 (86.2%)	1.74 (93.1%)
Reflect. posts	56 (113)	56 (self test)	2.97 (33.6%)	1.78 (59.8%)	1.85 (66.3%)
Train. set 1		6		2.10 (51.3%)	2.25 (52.2%)
Train. set 2		7		2.20 (51.3%)	2.28 (51.3%)
Train. set 3		7		2.07 (51.3%)	2.36 (52.2%)
Train. set 4		5		2.10 (51.3%)	2.30 (51.3%)

Table 1
Linear combination of BU and TD saliency, influence on search performance ($\lambda = 0$ equals pure BU and $\lambda = 1$ pure TD search)

sub-system during object search. Moreover, these examples visualize our understanding of the attention system as a common tunable front-end for the various other system tasks, e.g., for lane marking detection (see [39] for details on how the attention system can be used for lane marking detection). Following this concept, the task-specific tunable attention system can be used for scene decomposition and analysis, as it is shown exemplarily on two typical German highway scenes in Fig. 10.

Comparability of modalities: The used dynamic adaptation of w_{C_j} causes a twofold performance gain. First, the a-priori incomparable modalities get comparable yielding a well balanced BU and TD saliency map. Secondly, the system adapts to the dynamics of the environment preventing varying modalities from influencing the system performance (e.g., in the red evening sun the color R channel will not be overrepresented in the saliency). Figure 12b depicts the dynamically adapted w_{C_j}. Table 2 shows a noticeable gain in the object's signal to noise ratio $\overline{\text{SNR}}_{obj}$ on the overall saliency for 26 traffic relevant objects of an inner-city stream (see Fig. 11d), comparing the dynamically adapted w_{C_j} with static, locally optimized w_{C_j} vector. More specifically, $\overline{\text{SNR}}_{obj}$ is defined as the ratio of the mean saliency activation within the object region to the mean saliency activation in its surround. For system evaluation we set the smoothing parameter α (see Eq. (9)) to 0.05 in order to get a slow modality adaptation suitable to the moderate vehicle speed in the used inner-city stream.

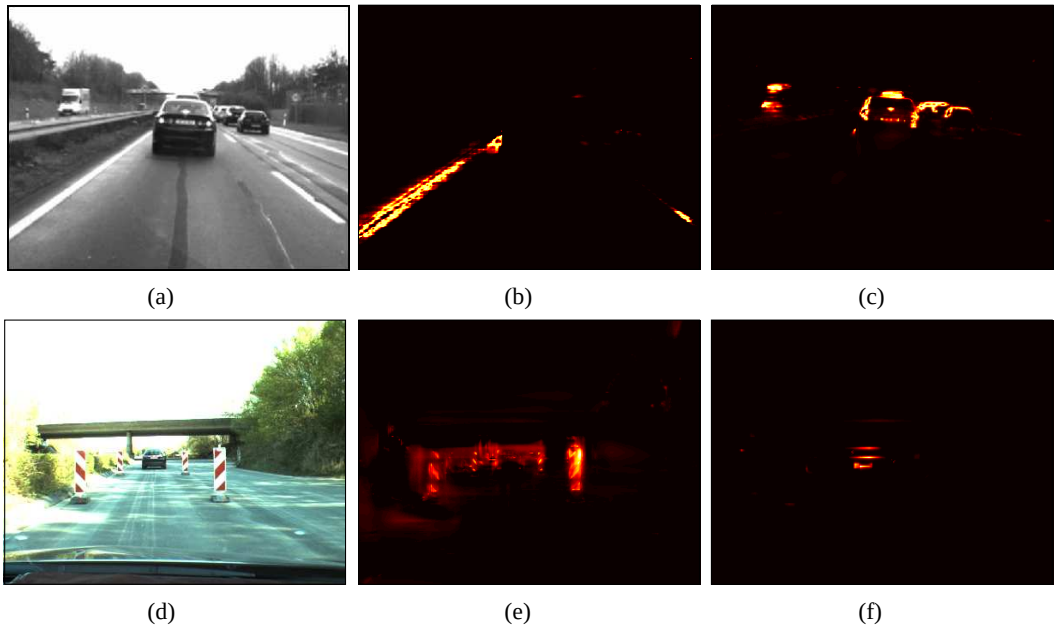


Fig. 10. Attention-based scene decomposition: (a) Highway scene, (b) TD attention tuned to lane markings, (c) TD attention tuned to cars, (d) Construction site, (e) TD attention tuned to signal boards, (f) TD attention tuned to cars.

Traffic-relevant objects	#images (objects)	$\overline{\text{SNR}}_{obj}$ using static w_{C_j}	$\overline{\text{SNR}}_{obj}$ using dynamic w_{C_j}
Inner city stream	20 (26)	2.56	2.86 (+11.7%)

Table 2
Comparability of modalities via homeostasis.

Evaluation of Classifier performance: For a proof of concept, we trained the classifier to distinguish cars from non-cars (clutter). A set of image segments generated by our vision system during online operation was used for training. It contains 11000 roughly square image patches scaled to a size of 64x64 pixels, and was divided into the classes 'car' (2952 patches), 'signal boards' (2408 patches) and 'clutter' (5803 patches) by visual inspection. Car segments contain complete back-views of cars (at any position) which must be at least half as large as the patch in both dimensions. At equal false positive and true negative rates, for cars an error of 2.8 % and for signal boards an error of 7.1 % was obtained on an equally large test. The performance of the trained classifier is shown in form of a ROC (Receiver-Operator-Characteristic) curve that visualizes the trade-off between false positive (clutter recognized as objects) and false negative (objects recognized as clutter) detections when varying the classification thresholds (see Fig. 12c). The ROC was generated using 5-fold cross validation.

Evaluation of depth fusion: Figure 13 shows the EKF-based fusion of

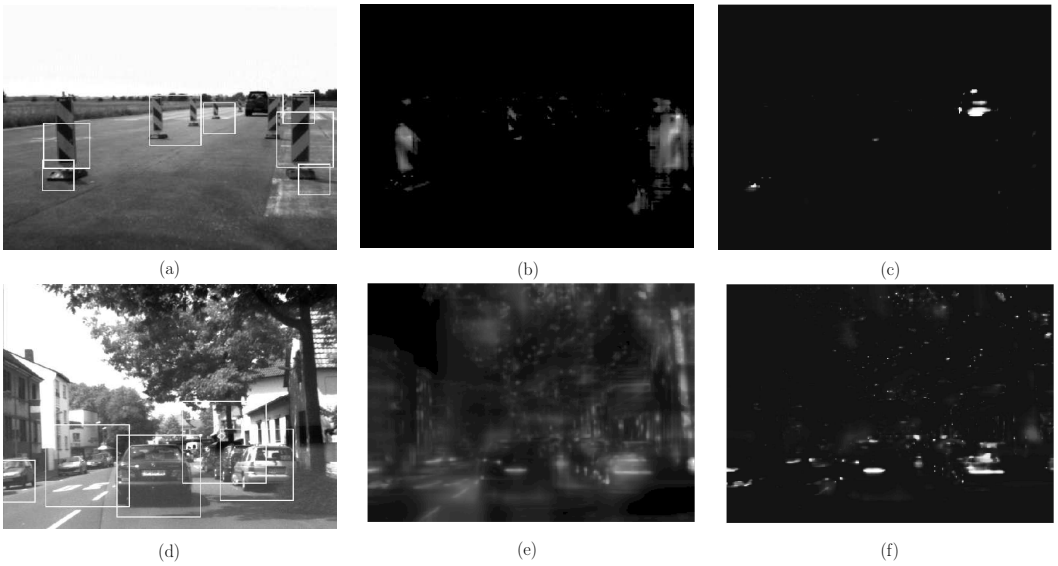


Fig. 11. Output of attention system for construction site and inner-city streams, (a) Unsegmented FoAs, tuned to signal boards, (b) TD saliency signal boards, (c) TD saliency cars, (d) Unsegmented FoAs, tuned to cars, (e) BU saliency, (f) TD saliency cars.

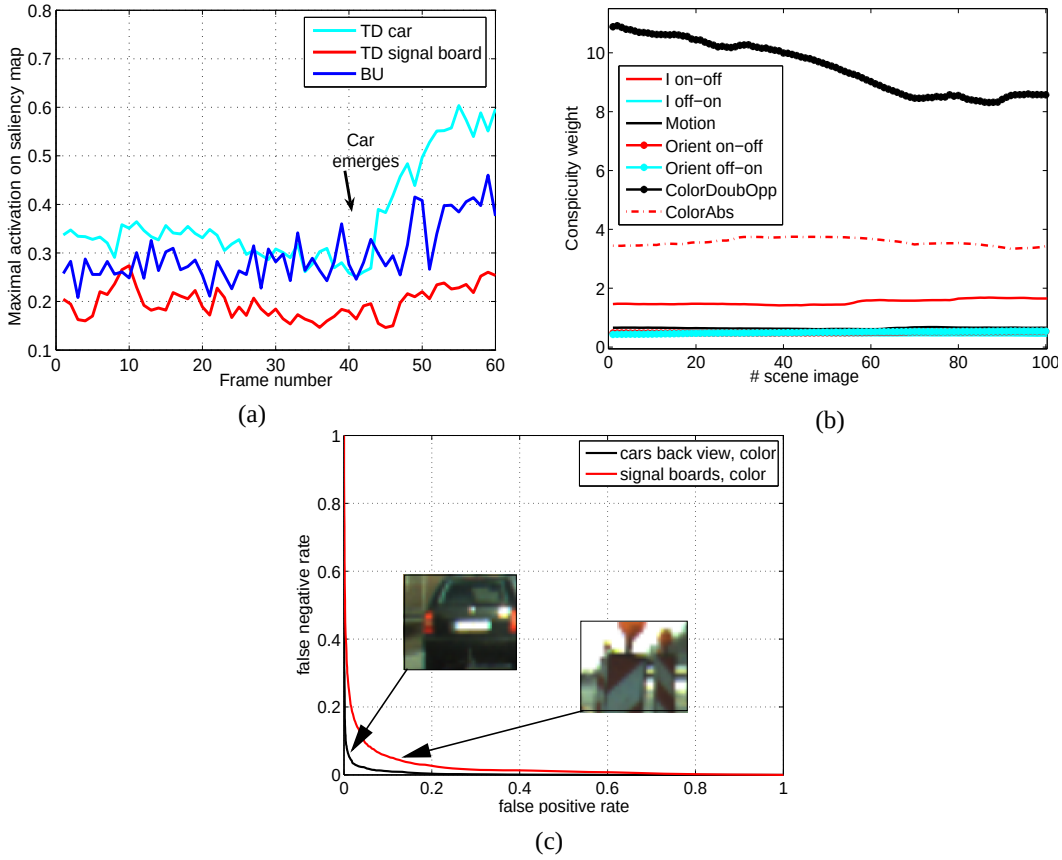


Fig. 12. (a) Normalization of features preserving magnitude information, (b) Comparability of modalities, (c) ROC curve for cars and signal boards.

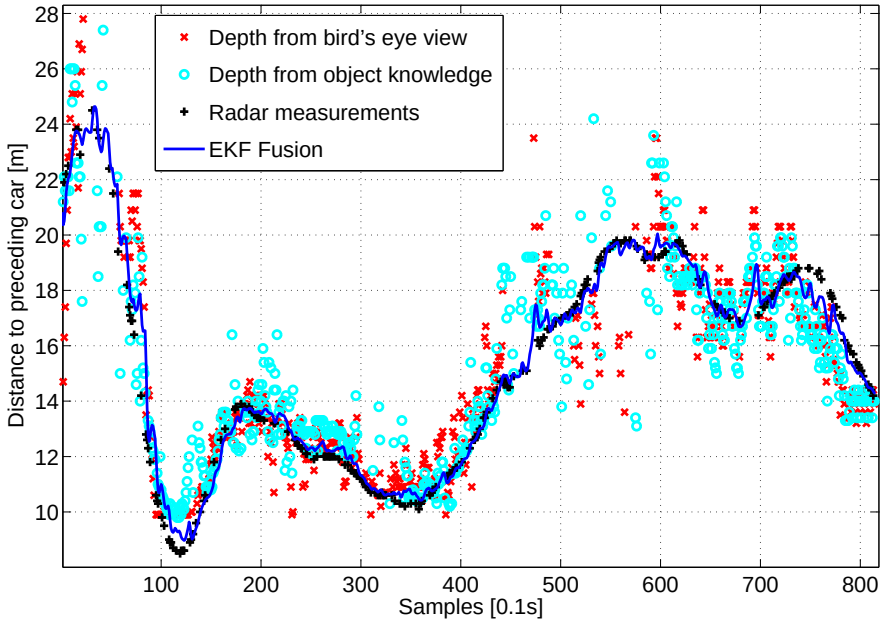


Fig. 13. Depth from bird's eye view, object knowledge, radar and fusion with EKF.

depth measurements from bird's eye view, object knowledge, and radar for a car that drives in front of our prototype vehicle through an inner-city (see Fig. 11d). The usage of two additional monocular depth cues of high variance fused with the low variance radar cue ensures the availability of depth cues even for objects that are far away. Here correctly located radar measurements seldom exist. For the EKF we used the static sensor variances $\sigma_{birds} = 2.8$, $\sigma_{obj} = 2.7$, and $\sigma_{radar} = 0.3$ as well as the process variance $\sigma_{process} = 0.023$.

4.2 Experimental Setup for System Evaluation

Scenario: In order to evaluate the proposed system in a challenging situation, we concentrate on typical construction sites on highways. This situation is quite frequent and a traffic jam ending exactly within a construction site is a highly dangerous situation: Due to the S-curve in many construction sites, the driver will notice a braking or stopping car quite late, see Fig. 14a. Our ADAS implementation uses a 3 phase danger handling scheme depending on the distance and relative speed of a recognized obstacle. For example, when the ego vehicle drives around 40 km/h and an obstacle is detected in front at less than 33 meters, a visual and acoustic warning is issued and the brakes are prepared. In the second phase, the belt pretensioners are triggered and the brakes are engaged with a deceleration of 0.25 g followed by hard braking of 0.6 g in the third phase.

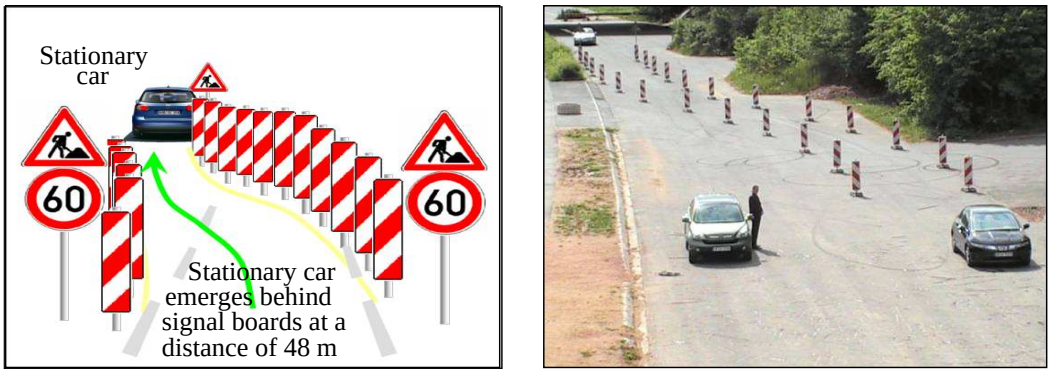


Fig. 14. Scenario: (a) Schematic sketch of the construction site scenario. Stationary car is visible from 48 meters on. (b) Real scenario.

Technical setup: For the experiments we use a Honda Legend prototype car equipped with a mvBlueFox CCD color camera from Matrix Vision delivering images of 800x600 pixels at 10Hz. The image data as well as the radar and vehicle state data from the CAN bus can be recorded. The recorded data is used during offline evaluation. For the online version, all data is transmitted via LAN to two Toshiba Tecra A7 (2 GHz Core Duo) running our RTBOS integration middleware [52] on top of Linux. The individual RTBOS components are implemented in C using an optimized image processing library based on the Intel IPP [53], allowing the overall system to run at 10Hz.

Test data for training and evaluation: In order to gain sufficient training data, we recorded image sequences during normal highway traffic including construction sites as well as visually complex scenes from driving in inner cities. For evaluating the actual system performance, we recorded data in an exemplary construction site on a private driving range.

4.3 Evaluation of System Performance

System statistics: During documented online system tests on our prototype vehicle (showing the setting depicted in Fig. 14) driving 40 km/h in 57 of 60 cases our system detected the stationary car in time and issued the 3 warning steps as expected including autonomous braking. In the remaining cases, either the object recognition detected a signal board as car and the braking was performed too early or the FoA generation did not deliver a good image segment of the stationary car so that the fusion of the image segment with radar data failed and no braking was performed at all.

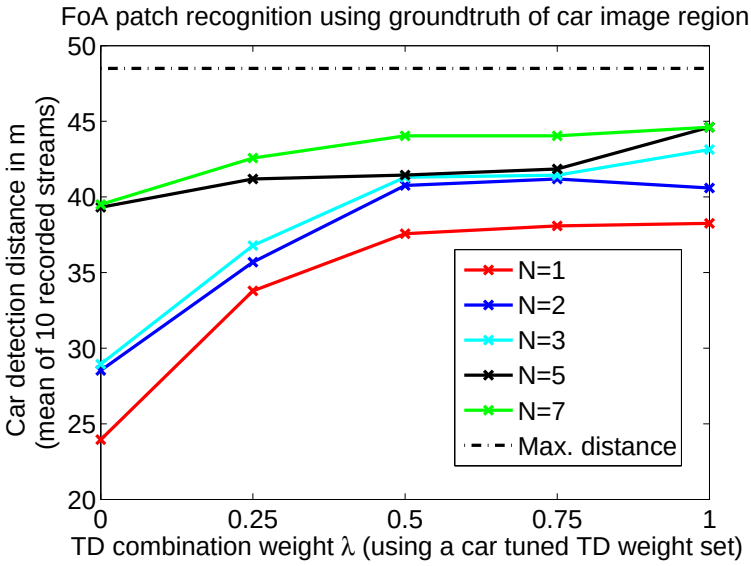
Furthermore we evaluated the warning generation offline in detail on 10 additional construction site streams we recorded. In all streams, the ADAS was able to recognize and track the car from a distance between 42 and 32 meters,

while the car was fully visible from a distance of about 48 meters on. On these recorded streams, we performed a more specific evaluation, described in the following.

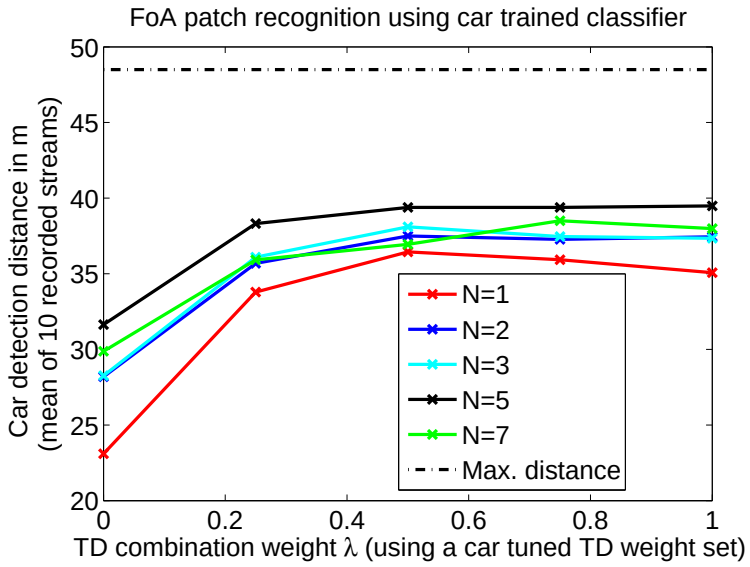
Influence of system parameters on the detection performance: In the following it is shown how the number of objects N stored in the STM influences the detection distance of the stationary car. Limiting the capacity of the STM in form of the parameter N is achieved by deleting an object from the STM after N frames. All depicted results are calculated by averaging over the 10 recorded streams in order to lessen statistical deviations. In the first step the car detection distance is evaluated depending on STM size N and the TD parameter λ (setting the amount of TD influence) while using a TD weight set trained on cars. Figure 15a shows the distance to the stationary car when the first FoA hits the car, which is defined by hand-labeled groundtruth on the recorded streams. It can be seen that the larger the TD influence (search task: find cars) expressed by λ, the earlier the car is detected. Similarly, the more objects are stored in the STM (object number N), the earlier the car is detected. It can also be deduced that with growing N the influence of TD is reduced since the scene coverage increases. Figure 15b shows the distance to the stationary car when the first FoA hits the target and the resulting image segment is recognized as car by the object classifier. Since the used classification threshold was set far above the equal false-positive false-negative error rate, the distance when the car is detected is smaller than in the evaluation with groundtruth. Differing from Fig. 15a, at a certain N the detection distance worsens again. The reason for this effect is that our system is not using crisp object segmentation algorithms but performs segmentation directly on the saliency map which can lead to enlarged patches suppressing the surround of the found objects as well. In this way, the borders of the car might be suppressed by adjacent signal board patches leading to incomplete car FoAs that are not sufficient for the used object classifier.

Based on Fig. 15 the best choice of λ would be 1, which equals pure TD search mode. Nevertheless such a parameterization is not appropriate as is shown in Fig. 16a. Here we see that with growing λ the average detection distance of signal boards drops. Stated differently, the system suffers from inattentional blindness while searching for cars in pure TD mode ($\lambda = 1$), which might lead to dangerous situations. The default value for λ was hence set to 0.5.

A parameter interacting with λ is the Time To Live [TTL] defining for how many frames an object is stored in the STM before it is removed. Figure 16b shows how the choice of the TTL influences the system performance. For an object-unspecific TTL of 5 frames the curve is identical to Fig. 16a for N=5. For the object-specific case we chose a TTL of 6 for signal boards, unknown objects were stored for 3 frames and cars for 20 frames, leading on average to N=5 objects in the STM. A clear gain in detection performance can be seen



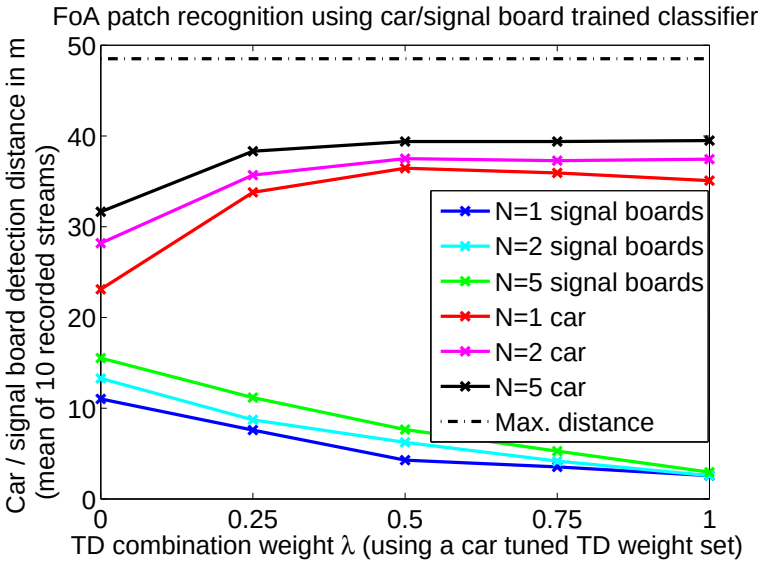
(a)



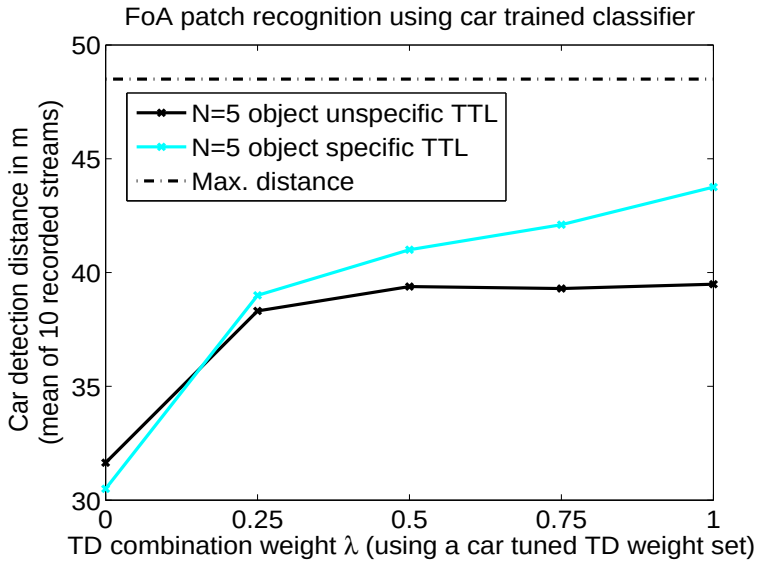
(b)

Fig. 15. Stationary car detection distance depending on λ=0, 0.25, 0.5, 0.75, and 1 as well as the STM size N=1,2,3,5, and 7. (a) Using groundtruth for detecting a hit, (b) Using the classifier for detecting a hit.

while using object-dependent TTL values which is due to the fact that FoAs, which hit the car very early are too small for a reliable classification. These unknown scene parts should not be suppressed too long in order to soon give the classifier a second chance to detect the car. The described object-specific TTL parameterization was also used during our online tests.



(a)



(b)

Fig. 16. Detection distance depending on λ=0, 0.25, 0.5, 0.75, and 1. (a) Average detection distance of signal boards and the stationary car using the object classifier for a STM size of N=1,2, and 5 (b) Stationary car detection distance while using object-specific TTL values.

5 CONCLUSIONS AND FUTURE WORKS

The contribution introduced an integrated vision architecture for ADAS, which realizes cognitive principles. Encouraging results obtained from the application of an attention system that can be modulated in a task-oriented top-down style were presented. The system is working online performing an autonomous

braking functionality on a Honda Legend prototype car. Our future work will concentrate on the online adaptation and unsupervised training of TD weight sets. We plan, to readapt the trained TD weight sets constantly depending on changes in the environment and a temporal decay. TD weight sets will be calculated offline for a limited number of traffic-relevant objects, like traffic signs and cars. Additionally, we plan to be not bound to this pre-calculated set and extract new TD weight sets online to track and find objects stored in the STM.

As an extension to the here described system, the ADAS presented in [39] contains an internal 3D representation, an unmarked road recognition system, broader information fusion, as well as a computational attention system that allow the online calculation of TD weights and thereby the simultaneous search for different object classes. We currently port the there described extensions from Matlab to C in order to integrate them in our existing online system [54] for evaluating them on our prototype vehicle. After the successful test of the low complexity control approach, in the next step, learning of the functional mapping between the measured input feature space and the output control parameter space will be in our focus. More specifically, we plan to replay stored streams of critical traffic situations from a data base. As learn-signal dangerous objects will be manually labeled in these streams. The system task is to detect the objects early enough. In case the system is too slow, the scenario is replayed by the learning algorithm while changing the functional mapping between input and output data of the behavior control module. Also measuring and mimicking the reactions of an experienced driver is envisioned in the future. Our system extensions introduced in [39] and [42] contains first approaches towards such an efficient cognitive control concept. The central assumption is that a robust learning system requires a generic system structure with a high number of degrees of freedom for controlling the system reaction and measuring the system state. Therefore, if required, we plan to further increase the input feature space as well as output control parameter space of the there described behavior control module in order to increase the number of possible system behaviors and prove the scalability of this extended approach.

6 ACKNOWLEDGMENTS

The authors gratefully acknowledge the support from Sven Bone, Falko Waibel, and Dr. Jens Gayko from Automobile Advanced Technology Research, Honda R & D Europe, for obtaining training data and demonstrating the system online on a prototype car. Also numerous reviewers' comments improved this contribution significantly.

References

[1] S. Frintrop, Vocus: A visual attention system for object detection and goal-directed search, Ph.D. thesis, University of Bonn Germany (2006).

[2] H. I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems: Sampling the Spectrum of Approaches, LNCS, Springer-Verlag, 2006.

[3] S. Treue, Visual attention: the where, what, how and why of saliency., in: Current Opinion in Neurobiology, Vol. 13, 2003.

[4] M. Ikegaya, N. Asanuma, S. Ishida, S. Kondo, Development of a lane following assistance system, in: Int. Symp. on Advanced Vehicle Control, Nagoya, 1998.

[5] S. Palmer, Vision Science: Photons to Phenomenology, MIT Press, 1999.

[6] M. Corbetta, G. Shulman, Control of goal-directed and stimulus-driven attention in the brain, Nature Reviews Neuroscience 3 (2002) 201–215.

[7] H. Egeth, S. Yantis, Visual attention: control, representation, and time course, Annual Review of Psychology 48 (1997) 269–297.

[8] D. Simons, C. Chabris, Gorillas in our midst: Sustained inattentional blindness for dynamic events, British Journal of Developmental Psychology 13 (1995) 113–142.

[9] J. M. Wolfe, T. S. Horowitz, What attributes guide the deployment of visual attention and how do they do it?, Nature Reviews Neuroscience 5 (6) (2004) 495–501.

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

[11] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[12] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artificial Intelligence 78 (1-2) (1995) 507–545.

[13] V. Navalpakkam, L. Itti (Eds.), A Goal Oriented Attention Guidance Model, Springer-Verlag, 2002.

[14] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vision Research 45 (2) (2005) 205–231.

[15] S. Frintrop, G. Backer, E. Rome, Goal-directed search with a top-down modulated computational attention system, in: Lecture Notes in Computer Science, 2005, pp. 117–124.

[16] N. Hawes, J. Wyatt, Towards context-sensitive visual attention, in: Proceedings of the Second Int. Cognitive Vision Workshop, Graz, Austria, 2006.

[17] C. Goerick, H. Wersing, I. Mikhailova, M. Dunn, Peripersonal space and object recognition for humanoids, in: Proc. Int. Conf. on Humanoid Robots, 2005.

[18] Z. Aziz, B. Mertsching, Visual search in static and dynamic scenes using fine-grain top-down visual attention, in: Lecture Notes in Computer Science, Vol. 5008, 2008, pp. 3–12.

[19] G. Backer, Modellierung visueller Aufmerksamkeit im Computer-Sehen: Ein zweistufiges Selektionsmodell fur ein Aktives Sehsystem, Ph.D. thesis, University of Hamburg Germany (2004).

[20] F. H. Hamker, The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision, Vol. 100, 2005, pp. 64–106.

[21] V. Navalpakkam, L. Itti, An integrated model of top-down and bottom-up attention for optimal object detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, 2006, pp. 2049–2056.

[22] J. Tsotsos, M. Pomplun, J. Martinez-Trujillo, K. Zhou, Attending to visual motion: Localizing and classifying affine motion patterns, in: CRV '04: Proceedings of the 1st Canadian Conference on Computer and Robot Vision (CRV'04), IEEE Computer Society, Washington, DC, USA, 2004, pp. 452–462.

[23] G. Backer, B. Mertsching, Integrating depth and motion into the attentional control of an active vision system, in: G. Baratoff, H. Neumann, (Eds.), Dynamische Perzeption, St. Augustin (Infix), 2000, pp. 69–74.

[24] T. Michalke, J. Fritsch, C. Goerick, Enhancing robustness of a saliency-based attention system for driver assistance, in: The 6th International Conference on Computer Vision Systems (ICVS), Santorini, Greece, 2008. Lecture Notes in Computer Science, Springer, No. 5008, 2008, pp. 43–55.

[25] J. Findlay, I. Gilchrist, Active Vision: The psychology of looking and seeing, Oxford University Press, 2003.

[26] S. Most, R. Astur, Feature-based attentional set as a cause of traffic accidents, Visual Cognition 15 (2007) 125–132.

[27] H. Shinoda, M. M. Hayhoe, A. Shrivastava, What controls attention in natural environments, Vision Research (41) (2001) 3535 – 3546.

[28] N. Ouerhani, Visual attention: From bio-inspired modeling to real-time implementation, Ph.D. thesis, Université de Neuchâtel, Institute de Microtechnique (2003).

[29] E. Dickmanns, Three-Stage Visual Perception for Vertebrate-type Dynamic Machine Vision, in: Engineering of Intelligent Systems (EIS), Madeira, 2004.

[30] G. Färber, Biological aspects in technical sensor systems, in: Proc. Advanced Microsystems for Automotive Applications, Berlin, 2005, pp. 3–22.

[31] C. Stiller, G. Färber, S. Kammel, Cooperative cognitive automobiles, in: IEEE Intelligent Vehicles Symposium, 2007, pp. 215–220.

[32] S. Matzka, Y. Petillot, A. Wallace, Proactive sensor-resource allocation using optical sensors, in: VDI-Berichte 2038, 2008, pp. 159–167.

[33] T. Michalke, A. Gepperth, M. Schneider, J. Fritsch, C. Goerick, Towards a human-like vision system for resource-constrained intelligent cars, in: Int. Conf. on Computer Vision Systems, Bielefeld, 2007.

[34] European Project PReVENT (2006).
URL `http://www.prevent-ip.org/`

[35] P. Cavanagh, G. Alvarez, Tracking multiple targets with multifocal attention, Trends in Cognitive Sciences 9 (2005) 350–355.

[36] R. M. Klein, Inhibition of return, Trends in Cognitive Science 4 (4) (2000) 138–145.

[37] A. Denecke, H. Wersing, J. Steil, E. Koerner, Online figure-ground segmentation with adaptive metrics in generalized LVQ, Neurocomputing.

[38] A. Broggi, Robust real-time lane and road detection in critical shadow conditions, in: Proc. Int. Symp. on Computer Vision, IEEE, Parma, 1995.

[39] T. Michalke, R. Kastner, J. Adamy, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko, A. Gepperth, J. Fritsch, C. Goerick, An attention-based system approach for scene analysis in driver assistance, at - Automatisierungstechnik 56 (11) (2008) 575–584.

[40] R. Trapp, Stereoskopische korrespondenzbestimmung mit impliziter detektion von okklusionen, Ph.D. thesis, University of Paderborn Germany (1998).

[41] V. Navalpakkam, L. Itti, Optimal cue selection strategy, in: Advances in Neural Information Processing Systems, Vol. 19, MIT Press, Cambridge, MA, 2006, pp. 1–8.

[42] T. Michalke, R. Kastner, J. Fritsch, C. Goerick, Towards a proactive biologically-inspired advanced driver assistance system, in: IEEE Intelligent Vehicles Symposium, Xian, 2009.

[43] R. N. Hardy, Homeostasis, Arnold, 1983.

[44] H. Wersing, E. Körner, Learning optimized features for hierarchical models of invariant object recognition, Neural Computation 15 (2) (2003) 1559–1588.

[45] A. Gepperth, B. Mersch, C. Goerick, J. Fritsch, Color object recognition in real-world scenes, in: J. de Sa (Ed.), J. Marques de Sa et al. (Eds.): Artificial Neural Networks, 17th International Conference ICANN, Part II, Lecture Notes in computer science, Springer Verlag Berlin Heidelberg New York, 2007, pp. 583–592.

[46] U. Neisser, Cognitive Psychology, Appleton-Century-Crofts, New York, 1967.

[47] P. Viola, M. J. Jones, Robust real-time object detection, in: 2nd Intl Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing and Sampling, 2001.

[48] M. Landy, L. Maloney, E. Johnsten, M. Young, Measurement and modeling of depth cue combinations: in defense of weak fusion (1995).

[49] H. Mallot, Computational vision: Information processing in perception and visual be-havior, MIT Press Robotica, 2002.

[50] V. Willert, J. Eggert, J. Adamy, E. Koerner, Non-gaussian velocity distributions integrated over space, time and scales, IEEE Transactions on Systems, Man and Cybernetics B 36 (3) (2006) 482–493.

[51] M. S. von Trzebiatowski, A. Gern, U. Franke, U.-P. Kaeppeler, P. Levi, Detecting reflection posts - lane recognition on country roads, in: IEEE Intelligent Vehicles Sysposium, 2004, pp. 304 –309.

[52] A. Ceravola, F. Joublin, M. Dunn, J. Eggert, C. Goerick, Integrated research and development environment for real-time distributed embodied intelligent systems, in: Proc. Int. Conf. on Robots and Intelligent Systems, 2006, pp. 1631–1637.

[53] Intel, Integrated Performance Primitives, `http://www.intel.com/cd/software/products/asmo-na/eng/perflib/ipp/302910.htm` (2006).

[54] J. Fritsch, T. Michalke, A. Gepperth, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko, C. Goerick, Towards a human-like vision system for driver assistance., in: IEEE Intelligent Vehicles Symposium, Eindhoven, 2008.