An Attention-based System Approach for Scene Analysis in Driver Assistance

Ein aufmerksamkeitsbasierter Systemansatz zur Szenenanalyse in der Fahrerassistenz

Thomas Michalke, Robert Kastner, Jürgen Adamy, Sven Bone, Falko Waibel, Marcus Kleinehagenbrock, Jens Gayko, Alexander Gepperth, Jannik Fritsch, and Christian Goerick

Research on computer vision systems for driver assistance resulted in a variety of isolated approaches mainly performing very specialized tasks like, e. g., lane keeping or traffic sign detection. However, for a full understanding of generic traffic situations, integrated and flexible approaches are needed. We here present a highly integrated vision architecture for an advanced driver assistance system inspired by human cognitive principles. The system uses an attention system as the flexible and generic front-end for all visual processing, allowing a task-specific scene decomposition and search for known objects (based on a short term memory) as well as generic object classes (based on a long term memory). Knowledge fusion, e. g., between an internal 3D representation and a reliable road detection module improves the system performance. The system heavily relies on top-down links to modulate lower processing levels, resulting in a high system robustness.

Bildbasierte Fahrerassistenzsysteme verfügen in der Regel über starre Funktionen, die sehr spezialisierte Aufgaben, wie Spurhaltung oder Verkehrszeichenerkennung, in fest definierten Situationen bearbeiten. Fahrerassistenzsysteme, die in einer großen Bandbreite von möglichen Verkehrssituationen robust und sinnvoll reagieren sollen, benötigen jedoch integrierte und flexiblere Ansätze. In der vorliegenden Arbeit wird ein integriertes Fahrerassistenzsystem vorgestellt, dessen Bildverarbeitungssubsystem durch Signalverarbeitungsprozesse im menschlichen Gehirn motiviert ist. Das Subsystem verwendet ein biologisch motiviertes Aufmerksamkeitsmodul als flexibles und generisches Front-end für alle Bildverarbeitungsprozesse. Das Aufmerksamkeitsmodul erlaubt eine aufgabenabhängige Szenenzerlegung, das Wiederfinden von bereits erkannten Objekten aus dem Kurzzeitspeicher des Systems sowie die generische Detektion von beliebigen Objektklassen über den Langzeitspeicher des Systems. Die Fusion von Informationen verschiedener Teilmodule, z. B. zwischen der internen 3D-Umfeldrepräsentation und einem Modul zur Detektion von unmarkierten Straßenflächen, erhöht die Güte des Gesamtsystems. Der Ansatz verwendet rekurrente Signalwege (so genannte top-down Verbindungen), welche Module auf tieferen Systemstufen online dynamisch parametrisieren, um die Robustheit und Reaktionsgeschwindigkeit des Gesamtsystems zu verbessern.

Keywords: Attention, human-like signal processing, task-dependent scene interpretation

Schlagwörter: Aufmerksamkeit, menschliche Signalverarbeitung, aufgabenabhängige Szeneninterpretation

1 Introduction

The goal of realizing Advanced Driver Assistance Systems (ADAS) can be approached from two directions: either searching for the best engineering solution or taking the human as a role model. Today's ADAS are engineered for supporting the driver in clearly defined traffic situations like, e. g., keeping the distance to the forward vehicle. While it may be argued that the quality of an engineered system in terms of isolated aspects, e. g., object detection

or tracking, is often sound, the solutions lack necessary flexibility. Small changes in the task and/or environment often lead to the necessity of redesigning the whole system in order to add new features and modules, as well as adapting how they are linked. In contrast, biological vision systems are highly flexible and are capable of adapting to severe changes in the task and/or the environment. Hence, one of our design goals on our way to achieve an "all-situation" ADAS is to implement a biologically motivated, cognitive vision system as perceptual front-end of an ADAS, which can handle the wide variety of situations typically encountered when driving a car. Note that only if an ADAS vision system attends to the *relevant* surrounding traffic and obstacles, it will be fast enough to assist the driver in real time during all dangerous situations.

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., [21]). Consequently, a cognitive vision system should realize a task-dependent perception using top-down links for modulating and parameterizing submodules, that is operating successfully without being explicitly designed for specific tasks of a scenario. Using this paradigm, the same scene can be decomposed by the vision system in different ways depending on the current task.

In order to realize such a cognitive vision system we have developed a robust attention sub-system [8] that can be modulated in a task-oriented way, i. e., based on the current context. The attention sub-system is a central component of the overall vision system realizing temporal organization of different visual processes. Its architecture is inspired by findings of human visual system research (see, e. g., [13]) and organizes the different functionalities in a similar way. In a first proof of concept, we have shown that a purely saliency-based attention generation can assist the driver during a critical situation in a construction site by performing autonomous braking [12].

While our earlier work concentrated mainly on saliency-based attention [8; 12], this contribution describes the additional incorporation of environmental 3D representations and static domain specific tasks, in order to use context information ("where is the road") to guide attention and, therefore, analysis of the overall scene. For all acquired information our enhanced system builds up internal 3D representations that support scene analysis and at the same time serve for behavior generation. Using a metric representation of the road area in combination with detected traffic objects, the system can guide its processing on relevant objects in the context of the current road area. For example, this allows to perform warning and emergency braking if a parked car is detected on our lane and during its by-passing the pro-actively adapted attention detects oncoming traffic on the road.

2 Related work

Recently, the topic of researching intelligent cars is gaining increasing interest as documented by the DARPA Urban Challenge [1] and the European Information Society 2010 *Intelligent Car Initiative* [2] as well as several European Projects like, e. g., Safespot or PReVENT.

Regarding vision systems developed for ADAS, there have been few attempts to incorporate aspects of the human visual system into complete systems. In terms of complete vision systems, one of the most prominent examples is a system developed in the group of E. Dickmanns [3]. 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 top-down aspects that are limited to a number of object-specific features for classification, no dynamic preselection of image regions is performed. A more biologically inspired approach has been presented by Färber [4]. This publication as well as the recently started German Transregional Collaborative Research Centre 'Cognitive Automobiles' [5] address mainly human inspired behavior planning whereas our work currently focuses more on task-dependent perception aspects.

More specifically, in the center of our work is a computational model of the human attention system that determines the how and when of scene decomposition and interpretation. Attention is a principle that was found to play an important role in the human vision processing as a mediator between the world and our actual perception [6]. Somewhat simplified, the attention map shows high activation at image positions that are visually conspicuous, i. e., that pop out (bottom-up attention) or that are important for the current system task (top-down attention). Derived from the first computational attention model [17], which showed only bottom-up aspects, some more recent models have been developed that also incorporate top-down information (see, e. g., [7; 8; 18; 19]). Please refer to [8] for a comprehensive comparison between the state-of-the-art attention systems [7; 18] and our computational attention model.

Recently, some authors stress the role of incorporating context into the attention-based scene analysis. For example [20], proposes a combination of a bottom-up saliency map and a top-down context driven approach. The top-down path uses spatial statistics, which are learned during an offline learning phase, to modulate the bottom-up saliency map. This is different to the system described here, where no offline spatial prior learning phase is required. In our online system context is incorporated in the form of top-down weights that are modified at run time and road information, as will be described in Sects. 3.1 and 3.3.

To our knowledge in the car domain no task-dependent tunable vision system that mimics human attention processes exist.

3 System

The proposed overall architecture concept for a robust attention-based scene analysis is depicted in Fig. 1. It consists of four major parts: the "what" pathway, the "where" pathway, a part executing static domain specific tasks, and the behavior generation. The distinction between "what" and "where" processing path is somewhat similar to the human visual system where the dorsal and ventral pathway are typically associated with these two functions (see, e. g., [13]). Among other things, the "where" pathway in the human brain is believed to perform the localization and tracking of a small number of objects. In contrast, the "what" pathway considers the detailed analysis of a single spot in the image (see theories of spatial attention, e. g., spotlight theory [13]). Nevertheless, an ADAS also requires specific information of the road and its shape, generated by the static domain specific part.

3.1 The "what" pathway

Starting in the "what" pathway the 400×300 color input image is analyzed by calculating the attention map S^{total}. The attention map S^{total} results from a weighted linear combination of $N = 130$ biologically inspired input feature maps F_i (see Eq. (1)). More specifically, we filter the image using, among others Difference of Gaussian (DoG) and Gabor filter kernels that model the characteristics of neural receptive fields measured in the mammal brain. Furthermore, we use the RGBY color space [7] as attention feature that models the processing of photoreceptors on the retina. All features are computed on 5 scales relying on the well-known principle of image pyramids in order to allow computationally efficient filtering. All feature maps are postprocessed non-linearly in order to suppress noise and boost conspicuous or prominent scene parts (see [11] for a detailed description of these nonlinear processing steps).

The top-down (TD) attention can be tuned (i. e., parameterized) task-dependently to search for specific objects. This is done by applying a TD weight set w_i^{TD} that is computed and adapted online, based on Eq. (2), where the threshold $\phi = K_{conj} Max(F_i)$ with $K_{conj} = (0, 1]$ (see Fig. 2a for a visualization). The weights w_i^{TD} dynamically boost feature maps that are important for our current task/object class in focus and suppress the rest. The bottom-up (BU)

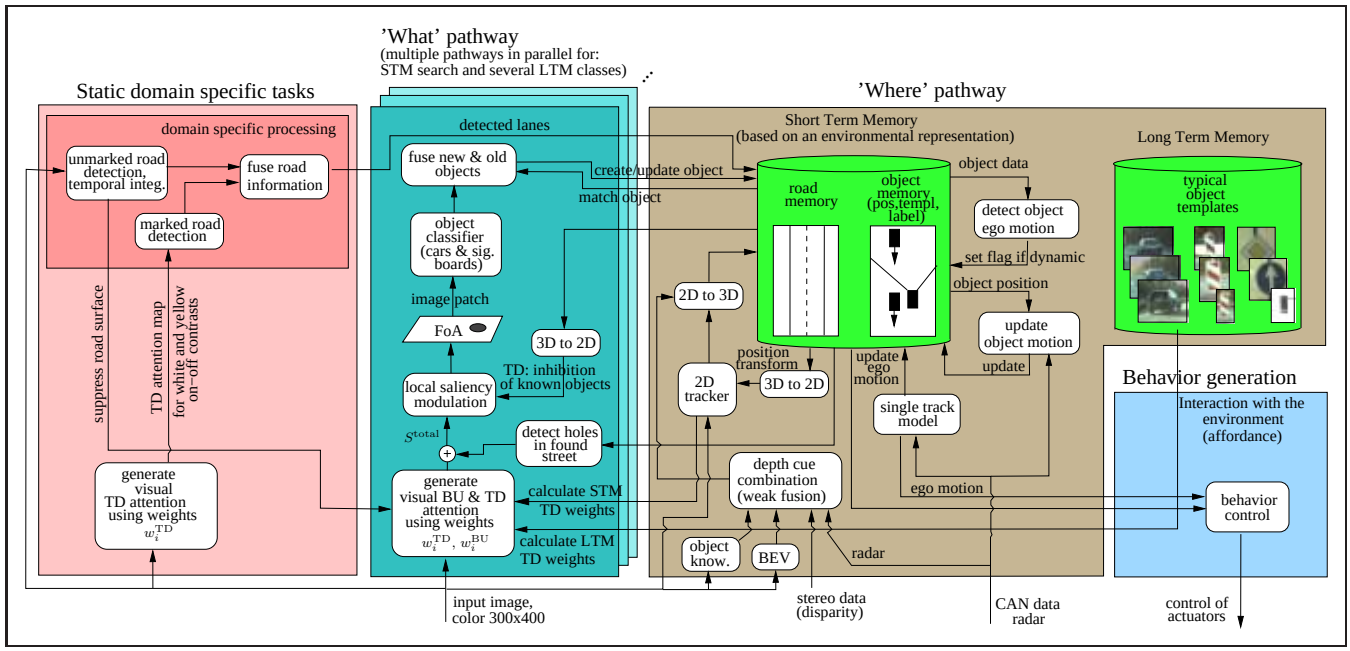

Figure 1: System structure allowing attention based scene analysis.

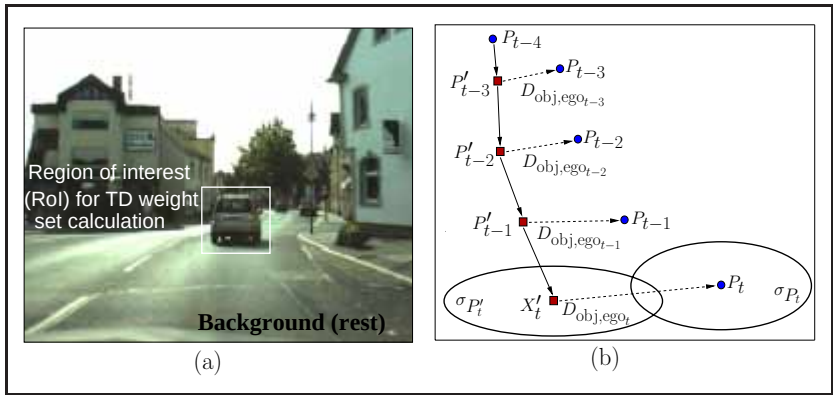

Figure 2: (a) Visualization of the object training region (RoI) for TD weight calculation against the background (rest), (b) Prediction of object ego motion (dots: Kalman tracked object position, squares: ego motion predicted object position, dashed line: accumulated object ego motion).

weights w_i^{BU} are set object-unspecifically in order to detect unexpected potentially dangerous scene elements. The parameter $\lambda \in [0, 1]$ (see Eq. (1)) determines the relative importance of TD and BU search in the current system state. For more details on the attention system please refer to [8]. It is important to note that the TD weights (calculated using Eq. (2)) are dependent on the features present in the background (rest) of the current image, since the background information is used to differentiate the searched object from the rest of the image [7]. Because of this, it is not sufficient to store the TD weight sets w_i^{TD} of different object classes directly and switch between them during online processing. Instead, all feature maps of objects $F_{i,RoI}$ are stored. To compensate the dependency from background the stored object feature maps are fused with the feature maps of the current image before calculating the TD weights. In plain words, the system takes the current scene characteristics (i. e., its features) into account in order to determine the optimal TD weight set that shows a maximum performance in the current frame. Put differently, the described separability approach includes the current scene context on a sensory level.

$$S^{total} = \lambda \sum_{i=1}^{N} w_i^{TD} F_i + (1-\lambda) \sum_{i=1}^{N} w_i^{BU} F_i \tag{1}$$

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

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

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

Now, we detect the maximum on the current attention map S^{total} and get the focus of attention (FoA) by generic region growing based segmentation on S^{total}. In the following, only the FoA is classified using a state-of-the-art object classifier that is based on neural nets [9]. This procedure (attention generation, FoA segmentation and classification) models the saccadic eye movements of mammals, where a complex scene is scanned and decomposed by sequential focusing of objects in the central 2–3° foveal retina area of the visual field. The system uses a time integrating mechanism to decide on the object class, in order to improve the reliability of the classifier decision. More specifically, all detected objects are tracked and reclassified in the following frames. On each frame a majority decision (voting) on the current and all stored classifier results decides on the object class.

The proposed system incorporates the biologically motivated concept of TD-links. Based on these links information on higher levels of knowledge integration modulate lower levels of knowledge integration. This brain-like concept improves robustness, increases the relevance of input data for higher system levels, and accelerates the system reaction (see evaluation results in Sect. 4). Our system uses such links for the task-specific modulation of the TD attention (i. e., by adapting system parameters online, as, e. g., the previously described TD-weights w_i^{TD}) and for suppressing the detected road (see Sect. 3.3) as context information in all feature maps F_i before fusing them in the overall saliency S^{total}. Additionally, TD-links are used for the modulation of the attention based on detected car-like holes in the found drivable road segment (see "where" path in Fig. 1). Such car-like holes are detected by searching for car-sized openings in the road memory, which is part of the 3D representation. Initially, the road segment is transformed to the metric bird's eye view (for an example see Fig. 4d) by inverse perspective mapping for the fusion with the 3D representation. In a nutshell, the bird's eye view is the representation of the scene as viewed from above, computed by transforming a monocular camera image taking intrinsic and extrinsic camera parameters into account (refer to [11] for more details).

3.2 The "where" pathway

The next step is the fusion between the newly detected object and the already known ones. The result will be further processed in the 'where' pathway and stored in the short term memory (STM). The objects in the STM are then suppressed in the current calculated attention map to enable the system to focus on new objects. The principle of suppressing known objects was proved to exist in the human vision system as well and is termed inhibition of return (IoR), refer to [10] for details.

All known objects are tracked using a 2D tracker that is based on normalized cross correlation (NCC). The tracker gets its anchor (i. e., the 2D pixel position where the correlation based object search on the new image will be started) from a Kalman filter based prediction on the 3D representation taking the ego motion of the camera vehicle and tracked object into account. The predicted 3D position is transformed to 2D pixel positions (x,y) using a pin hole camera model that contains all intrinsic and extrinsic camera parameters (in detail these are the 3 camera angles θ_X, θ_Y, and θ_Z, the 3 translational camera offsets t_1, t_2, t_3, the horizontal and vertical principal point c_u and c_v, as well as the horizontal and vertical focal length f_u and f_v), refer to Eqs. (3) and (4).

In case the NCC tracker is able to re-detect the object in 2D pixel coordinates, the 3D position in the representation is updated using 4 different depth cues for the 2D pixel (x, y) to 3D world $(X_{obj}, Y_{obj}, Z_{obj})$ transformation. More specifically, our system uses stereo data, radar, depth from object knowledge, and depth from bird's eye view (see Fig. 4 and [11; 12] for more details on these cues). The available depth cues are combined using the biologically motivated principle of weak fusion (see [16]). Weak

fusion combines the depth sources based on their reliability (i. e., sensor variances). The fusion is realized using an Extended Kalman Filter (EKF) that combines the cues based on dynamically adapted weights depending on the static predefined sensor variances and the available depth sources, 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 of the car (velocity and acceleration).

$$x = -f_u \frac{r_{11}(X\text{-}t_1) + r_{12}(Y\text{-}t_2) + r_{13}(Z\text{-}t_3)}{r_{31}(X\text{-}t_1) + r_{32}(Y\text{-}t_2) + r_{33}(Z\text{-}t_3)} + c_u \tag{3}$$

$$y = -f_v \frac{r_{21}(X\text{-}t_1) + r_{22}(Y\text{-}t_2) + r_{23}(Z\text{-}t_3)}{r_{31}(X\text{-}t_1) + r_{32}(Y\text{-}t_2) + r_{33}(Z\text{-}t_3)} + c_v \tag{4}$$

with

$$\begin{aligned}
r_{11} &= cos(\theta_Z)cos(\theta_Y)\\
r_{12} &= -sin(\theta_Z)cos(\theta_X) + cos(\theta_Z)sin(\theta_Y)sin(\theta_X)\\
r_{13} &= sin(\theta_Z)sin(\theta_X) + cos(\theta_Z)sin(\theta_Y)cos(\theta_X)\\
r_{21} &= sin(\theta_Z)cos(\theta_Y)\\
r_{22} &= cos(\theta_Z)cos(\theta_X) + sin(\theta_Z)sin(\theta_Y)sin(\theta_X)\\
r_{23} &= -cos(\theta_Z)sin(\theta_X) + sin(\theta_Z)sin(\theta_Y)cos(\theta_X)\\
r_{31} &= -sin(\theta_Y)\\
r_{32} &= cos(\theta_Y)sin(\theta_X)\\
r_{33} &= cos(\theta_Y)cos(\theta_X)
\end{aligned}$$

Objects whose updated position leave the represented surrounding scene or whose Kalman variances are too high (i. e., they received no new measurements for several frames) are deleted from the STM. The concept of appearance based 2D tracking (analysis of motion in 2D) supported by a 3D representation (interpretation of motion in 3D) was found in humans as well [13]. From a technical point of view, the advantage of this approach is the simple correction of the ego motion relying on the internal 3D representation. The vehicle ego motion (translations ΔX_e and ΔZ_e, as well as the change of the yaw angle $\Delta\theta_X$) is determined based on a standard single track model and compensated in the Kalman prediction step (see Eqs. (5) and (6) for the state vector E and process model A). Therefore, we do not need a computationally intensive optical flow based prediction. The main reason for the strong object motion in the 2D image is compensated by correcting the ego motion based position change of objects, which eases the tracking task considerably.

$$E = \left[Z_{obj} \;\; X_{obj} \;\; v_{Z,obj} \;\; v_{X,obj} \right] \tag{5}$$

$$A = \begin{bmatrix}
cos(\Delta\theta_X) & sin(\Delta\theta_X) & T & 0 & -\Delta Z_e \\
-sin(\Delta\theta_X) & cos(\Delta\theta_X) & T & 0 & -\Delta X_e \\
0 & 0 & cos(\Delta\theta_X) & sin(\Delta\theta_X) & 0 \\
0 & 0 & -sin(\Delta\theta_X) & cos(\Delta\theta_X) & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix} \tag{6}$$

A comparison between the current Kalman fused 3D object position $P_t = [Z_{obj}, X_{obj}]$ and the predicted 3D object position P'_t decides, based on the state variances $\sigma^2_{P'_t}$ and $\sigma^2_{P_t}$, if the tracked object is static or dynamic (see Fig. 2b). P'_t is calculated by an ego motion based prediction starting from the stored Kalman fused value P_{t-4}. For the comparison, β_{th} is used as a threshold on the measure $\beta(P_t, P'_t)$ defined in Eq. (7). The calculated measure is motivated from a statistical parameter test that checks for the equality of two distributions. It showed good performance on various test streams. If $\beta(P_t, P'_t)$ is bigger than β_{th} (i. e., the object is detected to be dynamic) the Kalman filter receives the object ego motion $v_{Z,obj} \neq 0$ and $v_{X,obj} \neq 0$ that is derived from the integrated object position change D_{obj,ego_t} as measurement.

$$\beta(P_t, P'_t) = \left| \frac{P_t - P'_t}{\sqrt{\sigma^2_{P_t} + \sigma^2_{P'_t}}} \right| \tag{7}$$

From a representational point of view, the "where" pathway of our system consists on the one hand of the STM, that stores all properties of sensed objects in a 3D representation and on the other hand of a long term memory (LTM) that stores the generic properties of object classes. The LTM is filled offline with typical patches and corresponding feature maps F_i of specific object classes. For evaluation purposes we use cars, reflection posts, and signal boards as LTM content, but our system can detect any other object types as well, if the attention and the object recognizer are trained accordingly. In the default state the system searches for the generic LTM object class car. This is done by calculating the geometric mean of all TD weight sets of the LTM objects that were calculated based on Eq. (2). These weights tune the TD attention in the "what" pathway.

As described above, in case the tracker has re-detected the object in the current frame the 3D representation is updated. In case the tracker looses the object the system searches for the lost STM object in the following frames. This is realized by calculating a TD weight set that is specific to the lost STM object using Eq. (2). The object O_f found by the STM search is then compared to the searched object O_s by means of the distance measure $\delta(O_f, O_s)$ that is based on the Bhattacharya coefficient (a measure for determining the similarity between two histograms) calculated on the histograms of all N object feature maps $H_i^{O_f}$ and $H_i^{O_s}$ (see Eq. (8)).

$$\delta(O_f, O_s) = \sum_{i=1}^{N} \sqrt{1 - \gamma(H_i^{O_f}, H_i^{O_s})} \tag{8}$$

$$\gamma\left(H_i^{O_f}, H_i^{O_s}\right) = \sum_{\forall x,y} \sqrt{H_i^{O_f}(x,y) H_i^{O_s}(x,y)}$$

The LTM and STM object search run in parallel as indicated visually in Fig. 1. It is important to note that our system is not restricted to the detection and tracking of cars, reflection posts, and signal boards. By using different LTM object patches and by offline training of our object

classifier in combination with the generic concept of online tunable TD attention our system is highly dynamic and flexible.

3.3 Static domain specific tasks

The third major part of our system handles the domain specific tasks of marked and unmarked road detection. The marked road detection is based on a standard Hough transform whose input signal is generated by our generic attention system. The scale-selective TD attention weight set used here boosts white and yellow structures on a darker background (so called on-off contrast), to which the biological motivated DoG filter (see Sect. 3.1) is selective. The yellow on-off structures are weighted stronger than the white to allow the handling of lane markings in construction sites.

The state-of-the-art unmarked road detection evaluates a street training region in front of the car and two non-street training regions at the side of the road. The features (stereo, edge density, color hue, color saturation) in the street training region are used to detect the drivable road based on dynamic probability distributions for all cues. Additionally, region growing that starts at the street training region assures a crisp distinction between the road and the sidewalk. The region growing uses dynamic self-adaptive thresholds that are derived from the feature characteristics in the street training as compared to the non-street training region. No fixed parameters for detecting the road are used, which makes the system adaptive to its environment and hence robust. A temporal integration procedure between the current and past detected road segments based on the bird's eye view is used to increase the completeness of the detected road by decreasing the number of false negative road pixels. More specifically, based on the measured egomotion of the car the road segments detected in the past are shifted and fused with the currently detected road segments. Refer to [11] for a comprehensive description of the temporal integration procedure. In the final step, a fusion between the marked and unmarked detected road segments is used to derive the present drivable lanes.

3.4 Behavior control

The system can interact with the world via a behavior control module. Currently our ADAS implementation uses a 3 phase danger handling scheme depending on the distance and relative speed of a recognized obstacle. When an obstacle is detected in front at a rapidly decreasing distance, a visual and acoustic warning is issued and the brakes are prepared. In the second phase the brakes are engaged with a deceleration of 0.25 g followed by hard braking of 0.6 g in the third phase (refer to [12] for details on the behavior control). Other behaviors, like trajectory planning and active steering, as well as the detection of possible collisions and their active avoidance based on predictions on internal 3D representations are possible and planned in the near future.

4 Results

In Sect. 4.1 we will evaluate different individual system modules that play an important role in our cognitive ADAS architecture. In Sect. 4.2 the overall system performance will be assessed based on a scenario showing a stationary car in a construction site.

4.1 Evaluation of system modules

Evaluation of attention subsystem: In order to evaluate the generic nature of the attention based TD search, we used cars and reflection posts (useful for unmarked road detection as done, e. g., in [14]) as LTM search objects. The results are depicted in Table 1, showing that incorporating TD information improves the search performance considerably. Please note that when changing the LTM search object, besides an exchange of the LTM image patches and an appropriate training of the object classifier no modification in the system structure is required. 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 $\overline{Hit}$ is, the earlier an object is detected (see [7] 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 show (see Table 1). The evaluation shows the highest hit numbers and detection rates for pure TD search ($\lambda = 1$). However, as will be discussed in Sect. 4.2 a combination of BU and TD influence in the attention system is recommended.

The presented results support the generic nature of the TD tunable attention subsystem during object search. Moreover, we see the attention system as a common tunable front-end for the various other system tasks, e. g., as lane

Table 1: Search performance for BU and TD based LTM object search for cars and reflection posts for 2 different training sets each.

Target	# Test images (objects)	# Training im	$\overline{Hit}$ ($\overline{DRate}$) pure BU ($\lambda = 0$)	pure TD ($\lambda = 1$)
Cars	54 (58)	54 (self test)	3.06 (56.9%)	1.53 (100%)
T.set 1		3		1.82 (96.6%)
T.set 2		3		1.74 (93.1%)
Reflect. posts	56 (113)	56 (self test)	2.97 (33.6%)	1.85 (66.3%)
T.set 1		6		2.25 (52.2%)
T.set 2		7		2.36 (52.2%)

marking detection (as described in Sect. 3.3). 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. 3.

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 11 000 square image patches of size 64×64 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 4.7% and for signal boards an error of 9.7% was obtained on equally large test sets. The performance of the trained classifier is shown in Fig. 7a in form of a ROC 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. The ROC was generated using 5-fold cross validation. Furthermore, the quality of the classification is enhanced by the voting process described in Sect. 3.1.

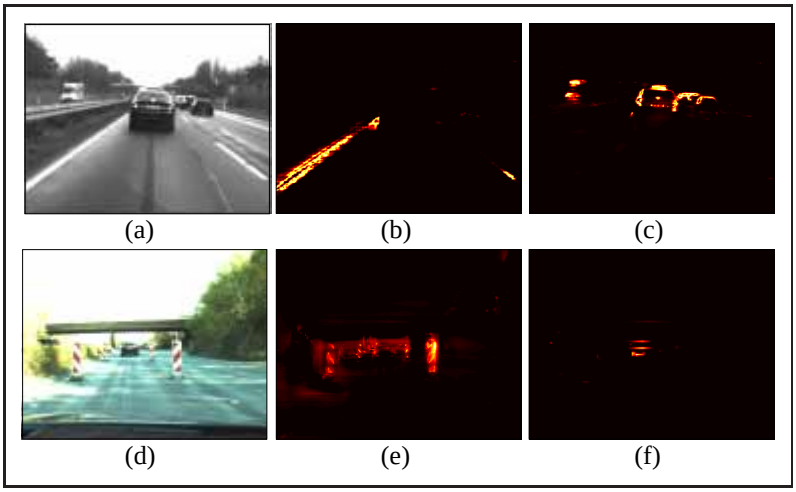

Figure 3: Attention based scene decomposition: (a) Highway scene, (b) TD attention tuned to lane markings, (c) TD attention tuned to cars, (d) Construction site (e) TD attention tuned to signal boards (f) TD attention tuned to cars.

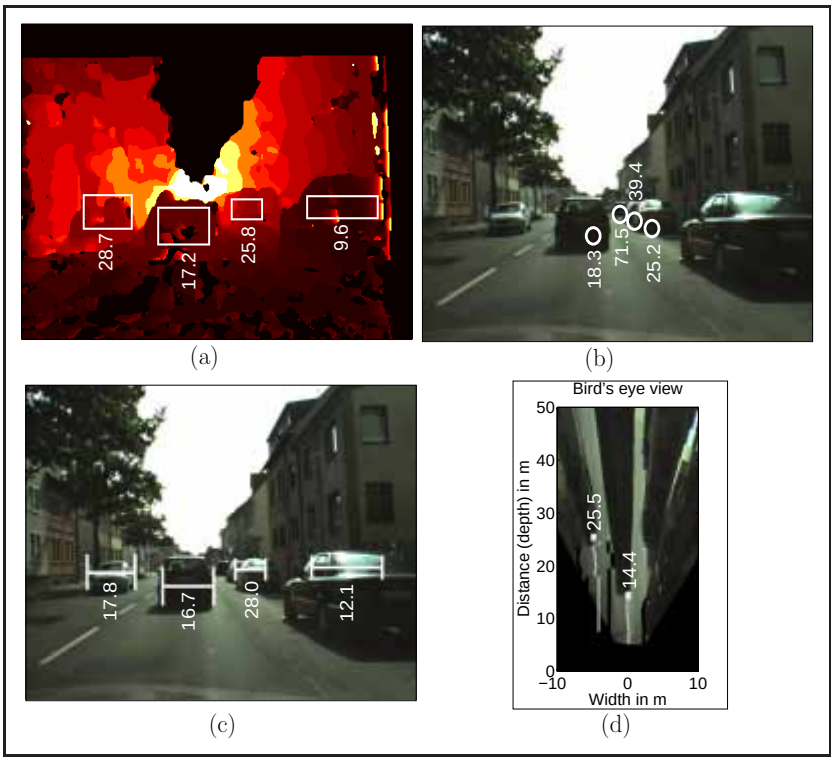

Figure 4: (a) Depth from Stereo (calculated as a median over the object region), (b) Depth from radar, (c) Depth from object knowledge (for all objects detected as cars), (d) Depth from bird's eye view (using threshold based detection of intensity changes on the road).

Qualitative evaluation of depth cues: For a more qualitative evaluation Fig. 4 shows the unpreprocessed results for all depth cues in a typical inner city stream. The cues show strong differences in accurateness (especially depth from bird's eye view and object knowledge show a high variance). However, this is uncritical, since the sensor variances (that were determined offline) are taken into account during the EKF based sensor fusion (see [12] for more a detailed depth cue evaluation).

Evaluation of unmarked lane detection: In order to evaluate how much of the pixels representing road area are classified correctly, we generated ground truth by manual hand labeling of 440 test images of an inner city stream. We use different ground truth based measures (see [15]) for evaluation (with pixels being True Positives (TP), False Negatives (FN), and False Positives (FP)):

$$\text{Completeness} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$\text{Correctness} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{Quality} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{11}$$

The Completeness states how much of the present street was really detected while the Correctness states how much of the detected street is actually street. The Quality combines both measures, as between the Completeness and Correctness a trade-off is possible (FN street pixels against FP street pixels). Therefore, the Quality should be used for comparison (it weights the FP and FN pixels equally) while Completeness and Correctness state what exactly caused a difference in Quality.

Table 2: Comparison of unmarked lane detection with and without temporal integration (TI).

Road detect. approaches	#test im.	Correct-ness	Complet-ness	Quality
Without TI	440	98.1%	61.5%	60.5%
With TI	440	**95.2%**	**94.1%**	**89.9%**

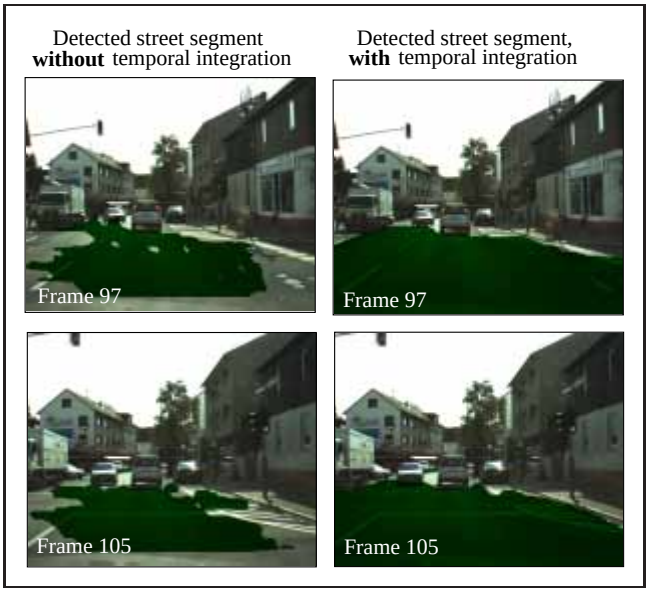

Figure 5: Road detection on example images of an inner city stream (left column: Without temporal integration, right column: With temporal integrated road segment).

Evaluation results for the road detection algorithm with and without temporal integration (see Sect. 3.3 and Table 2). A Quality of 89.9% is reached using temporal integration as opposed to a Quality of only 60.5% without temporal integration. While the highest Correctness of 98.1% is reached without temporal integration, this comes at the cost of a low Completeness and, consequently, a low Quality.

For further evaluation Fig. 5 shows the street detection results on a number of example frames of the inner city stream used for evaluation.

4.2 Evaluation of overall system performance

The performance gain of incorporating the detected drivable street, the internal metric 3D representation, and TD links are evaluated on a real-world construction site scenario. The results gathered with the proposed system are then compared with our previous system [12].

In a nutshell, the scenario described in [12] concentrated on typical construction sites on highways. 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. 6). The evaluation was done offline by averaging on 3 streams that were stored during the online demonstration of the previous ADAS. As depicted in Fig. 7b the current system architecture can classify the stationary car from 25 to 42 meters on. How early the car is detected depends on how much TD influence is incorporated. For $\lambda = 0$ the car is detected late, because only visually conspicuous object features are incorporated that draw BU attention. For a growing λ the car is detected early since "car-like" features are boosted stronger in the TD attention. Based on Fig. 7b the best choice of λ for detecting cars would be 1, which equals pure TD search mode. However, such a parameterization is not appropriate because this leads to a reduced capability of detecting other objects that are only prominent in the BU attention map. As depicted in Fig. 7b with growing λ (i. e., with growing influence of car features in the attention) the mean detection distance of signal boards as BU salient objects drops. Stated differently, the system ignores all other objects while searching for cars in pure TD mode ($\lambda = 1$), which might lead to dangerous situations. The measured effect was also proved to exist in biology and is termed "inattentional blindness" (see [6]). This suggests to set λ to an intermediate value of about 0.5, which was also the setting used during our online tests (see [12]).

Also compared to our previous system [12] for all λ values a better system performance was achieved. In our previ-

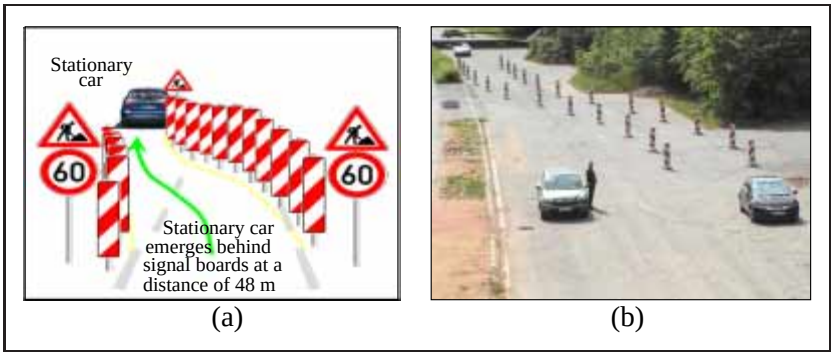

Figure 6: (a) Schematic sketch of the construction site scenario. Stationary car is visible from 48 meters on. (b) Real scenario.

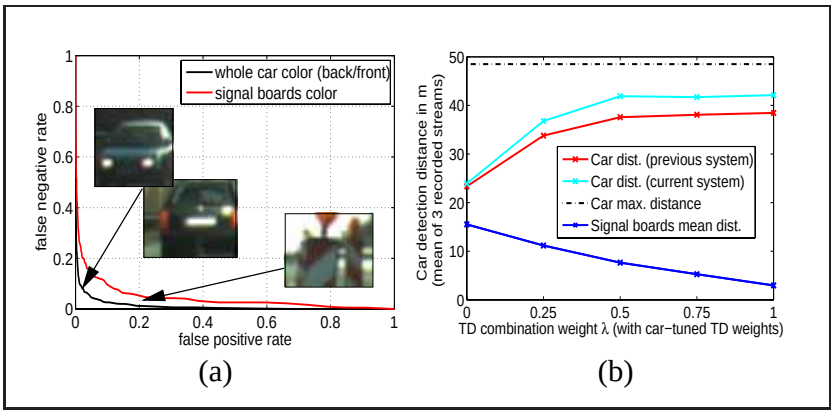

Figure 7: (a) ROC curve for cars (back and front views) and signal boards (b) Comparison between previous and current system implementation: Stationary car detection distance depending on $\lambda = 0$, 0.25, 0.5, 0.75, and 1. For both systems a comparable parameter set was used.

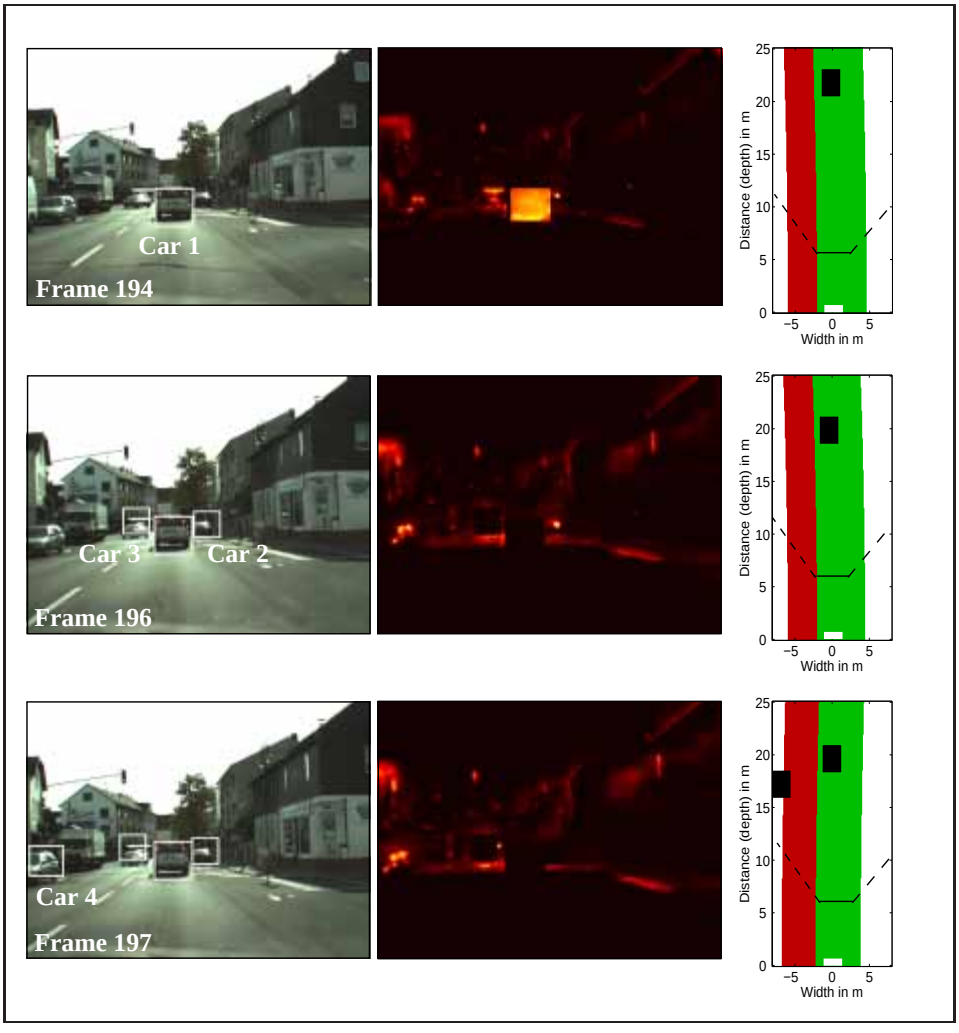

Figure 8: System evaluation on example images of an inner city stream (left column: visualization of found FoAs, middle: calculated attention map S^{total} (found objects are suppressed during IoR), right column: Visualization of internal representation (dashed line marks the border of the vision field)).

ous system an appearance based 2D tracking as opposed to the 3D tracking presented here was used. Furthermore, the TD weights were computed offline as opposed to the online LTM object search in the current system. Additionally, information drawn from the road detection module is included and combined to the attention module in the current system (see Sect. 3). The attained performance gain affirms the soundness of these cognitive system extensions.

For further system evaluation Fig. 8 depicts internal system variables for three sequential frames of an inner city stream with cars as LTM search object. As described in Sect. 3, for each new image the attention is calculated and a new FoA is generated via maximum search and segmentation on the attention map. The detected road area (and thereby also the present lane markings) are mapped out of the attention map, which decreases the false positive rate of generated FoAs, i. e., less non-car FoAs are generated. In the first frame, the car in front is detected and stored in the representation based on a car-like hole in the detected street segment that modulates the attention. Please note that car 2 and 3 are not stored in the internal representation, since their position is beyond the represented road environment.

5 Summary

In this contribution, we presented an integrated, advanced driver assistance system that relies on human-like cognitive processing principles. The system uses a biologically motivated attention system as the flexible and generic front-end for all visual processing. Based on TD links modulating the attention task-dependently, the used internal 3D representation, a state-of-the-art object classifier, and a road recognition system, we realized a highly flexible and robust system architecture. We currently port the described extensions from Matlab to integrate them in our existing online system [12], in order to evaluate them on our prototype vehicle.

References

[1] DARPA Urban Challenge. [Online]. Available: http://www.darpa.mil/ grandchallenge/.

[2] "European commission information society 'Intelligent Car initiative"', 2007. [Online]. Available: http://ec.europa.eu/informationsociety/activities/intelligentcar/.

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

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

[5] C. Stiller, G. Färber, and S. Kammel, "Cooperative cognitive automobiles", in IEEE Intelligent Vehicles Symposium, Istanbul, 2007, pp. 215–220.

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

[7] S. Frintrop: "VOCUS: A Visual Attention System for Object Detection and Goal- Directed Search", PhD thesis, University of Bonn Germany (2006).

[8] T. Michalke, J. Fritsch, and C. Goerick, "Enhancing Robustness of a Saliency-based Attention System for Driver Assistance", *Lecture Notes in Computer Science, Springer (5008), 43–55, 2008*.

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

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

[11] T. Michalke, R. Kastner, J. Fritsch, and C. Goerick, "A generic temporal integration approach for enhancing feature-based road-detection systems", *IEEE Intelligent Transportation Systems Conference, Beijing, 2008*.

[12] J. Fritsch, T. Michalke, A. Gepperth, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko, and C. Goerick, "Towards a Human-like Vision System for Driver Assistance", *IEEE Intelligent Vehicles Symposium, Eindhoven, 2008*.

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

[14] M. S. von Trzebiatowski, A. Gern, U. Franke, U.-P. Kaeppeler, P. Levi, "Detecting reflection posts - lane recognition on country roads", *IEEE Intelligent Vehicles Symposium, Parma, 2004*.

[15] P. Lombardi, M. Zanin, and S. Messelodi, "Unified stereovision for ground, road and obstacle detection", *IEEE Intelligent Vehicles Symposium, Las Vegas, 2005*.

[16] M. Landy, L. Maloney, E. Johnsten, M. Young, "Measurement and modeling of depth cue combinations: in defense of weak fusion", *Vision Research 35(3) (1995)*.

[17] C. Koch, S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry", *Human Neurobiology 4(4) (1985)*.

[18] V. Navalpakkam, L. Itti, "Modeling the influence of task on attention", *Vision Research 45(2) (2005)*.

[19] Z. Aziz, B. Mertsching, "Visual search in static and dynamic scenes using fine-grain top-down visual attention", *Lecture Notes in Computer Science, Springer (5008), 3–12, 2008*.

[20] A. Torralba, "Contextual priming for object detection", *International Journal of Computer Vision 53 (2003)*.

[21] S. Treue, "Visual attention: the where, what, how and why of saliency.", *Current Opinion in Neurobiology 13(4) (2003) 428–432*.

Manuscript received: 5th May 2008.

Dipl.-Wirtsch.-Ing. Thomas Michalke, Dipl.-Ing. Robert Kastner, Prof. Dr.-Ing. Jürgen Adamy

Address: Darmstadt University of Technology, Institute for Automatic Control, Control Theory and Robotics Lab, 64283 Darmstadt, E-Mail: {thomas.michalke, robert.kastner, adamy}@rtr.tu-darmstadt.de

Dipl.-Ing. Sven Bone, Dipl.-Ing. Falko Waibel, Dr.-Ing. Marcus Kleinehagenbrock, Dr.-Ing. Jens Gayko

Address: Honda Research & Development Europe GmbH, 63073 Offenbach/Main, E-Mail: {sven_bone, falko_waibel, marcus_kleinehagenbrock, jens_gayko}@de.hrdeu.com

Dr. rer. nat. Alexander Gepperth, Dr.-Ing. Jannik Fritsch, Dr.-Ing. Christian Goerick

Address: Honda Research Institute GmbH, 63073 Offenbach/Main, E-Mail: {alexander.gepperth, jannik.fritsch, christian.goerick}@honda-ri.de