Enhancing Robustness of a Saliency-based Attention System for Driver Assistance

Thomas Michalke, Jannik Fritsch, Christian Goerick

2008

Preprint:

This is an accepted article published in The 6th International Conference on Computer Vision Systems (ICVS). The final authenticated version is available online at: https://doi.org/[DOI not available]

Enhancing Robustness of a Saliency-based Attention System for Driver Assistance

Thomas Michalke[1], Jannik Fritsch[2], and Christian Goerick[2]

[1] Darmstadt University of Technology, Institute for Automatic Control
D-64283 Darmstadt, Germany
thomas.michalke@rtr.tu-darmstadt.de
[2] Honda Research Institute Europe GmbH, D-63073 Offenbach, Germany
{jannik.fritsch,christian.goerick}@honda-ri.de

Abstract. Biologically motivated attention systems prefilter the visual environment for scene elements that pop out most or match the current system task best. However, 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 or are tested in a controlled indoor environment only. No approach is so far used in the highly dynamic real world scenario car domain. Dealing with such scenarios requires a strong system adaptation capability with respect to changes in the environment. Here, we focus on five conceptual issues crucial for closing the gap between artificial and natural attention systems operating in the real world. We show the feasibility of our approach on vision data from the car domain. The described attention system is part of a biologically motivated advanced driver assistance system running in real time.

Keywords: driver assistance, top-down / bottom-up saliency, cognitive systems, real world robustness

1 INTRODUCTION

The most important sensory modality of humans with the highest information density is vision. The human vision system filters this high abundance of 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 mental status or need/task of the subject), while suppressing the rest. For both attention guiding principles psychophysical and neurological evidence exist (see [1, 2]). Following this principle, technical vision systems have been developed that prefilter a scene by decomposing it into basic features (see [3]) 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 [4]). More recent system implementations additionally include the modulatory influence of task relevance into the saliency (i.e., top-down (TD) attention, see [5] as one of the first and [6, 7] as the most recent and probably most influential approaches).

In these systems, instead of scanning the whole scene in search for 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 for 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.

In the following, we present a TD tunable attention system we developed that is the front end of the vision system of an advanced driver assistance system (ADAS) described in [8], whose architecture is inspired by the human brain. The design goals of our TD attention subsystem comprised the development of an object and task-specific tunable saliency suitable for the real world car domain. In this contribution we present new robustness enhancements in order to cope with the challenges our system is faced with when using saliency on real outdoor scenes. Important aspects discriminating real world scenes from artificial scenes are the dynamics in the environment (e.g., changing lighting and weather conditions, highly dynamic scene content) as well as the high scene complexity (e.g., cluttered scenes).

In Section 2 we will describe specific challenges that the mentioned aspects provoke on an attention system under real world conditions. Section 3 will describe our attention subsystem in detail pointing out the solutions to the denoted challenges and relates its structure to other attention approaches. Section 4 underlines the potential of the described solutions based on results calculated on different real world scenes after which the paper is summarized.

2 Real world challenges for TD attention systems

In the following paragraph we describe challenges a TD attention system is faced with when used on real world images.

① **High feature selectivity:** In order to yield high hit rates in TD search an attention system needs high feature selectivity to have as much supporting and inhibiting feature maps as possible. For this the used features must be selected and parameterized appropriately. Even more important for high selectivity is the use of modulatory TD weights on *all* subfeature maps and scales. Many TD attention approaches allow TD weighting only on a high integration level (e.g., no weighting on scale level [9]) or without using the full potential of features (e.g., no on-off/off-on feature separation [6]) which leads to a potential performance loss. Our system fulfills both aspects. Based on the extended selectivity of our attention subsystem, we can handle specific challenges of the car domain, as dealing with the horizon edge present in most images.

② **Comparable TD and BU saliency maps:** Typically the TD and BU attention maps are combined to an overall saliency, on which the Focus of Attention (FoA) is calculated. The combination requires comparable TD and BU

saliency maps, making a normalization necessary. Humans undergo the same challenge when elements popping out compete with task-relevant scene elements for attention. A prominent procedure in literature normalizes each feature map to its current maximum (see [6] that is based on [10]), which has some drawbacks our approach avoids.

③ **Comparability of modalities:** Similarly, the combination of different a priori incomparable modalities (e.g., decide on the relative importance of edges versus color) must be achieved. We realize this by the biological principle of homeostasis that we define as the reversible adaptation of essential processes of a (biological) system to the environment (see e.g., [11]).

④ **Support of conjunctions of weak object features in the TD path:** Another important robustness aspect is the support of conjunction of weak object features in the TD path of the attention subsystem. That is, an object having a number of mediocre feature activations but no feature map popping out should still yield a clear maximum when combined on the overall saliency.

⑤ **Changing lighting conditions:** In a real world scene changing lighting conditions influence the features the saliency is composed of and hence the attention system performance heavily. As the calculated TD weights are based on the features of the training images, the TD weights are illumination dependent as well. Put differently, the TD weights are optimal for the specific illumination and thereby contrast that is present in the training images. Using the TD weights on test images with a differing illumination will then lead to an inferior TD search performance. Therefore, a local exposure control is needed to adjust the contrast of the training images as well as the test images before applying TD weight calculation and TD search.

3 Modeling attention: From a robustness point of view

Section 3.1 focuses on the processing steps of our attention system that are crucial for solving the challenges described in Chapter 2. In Section 3.2 our system is compared to other state-of-the-art attention systems.

3.1 Description of the ADAS attention subsystem

The organization of Section 3.1 is led by the consecutive processing steps of the current ADAS attention subsystem as depicted in Fig. 1. After a short description of the general purpose of the BU and TD pathways, their combination to the overall saliency is described. Following this overview, the used modalities (feature types) are specified followed by the entropy measure that is used for the camera exposure control. Next, the different steps of the feature postprocessing are described. The TD feature weighting, the homeostasis process to get the conspicuity maps comparable, as well as the final BU/TD saliency normalization are the final processing steps in our attention architecture.

The attention system consists of a BU and a TD pathway. The TD pathway (on top) allows an object- and task-dependent filtering of the input data. All image regions containing features that match the current system task well are

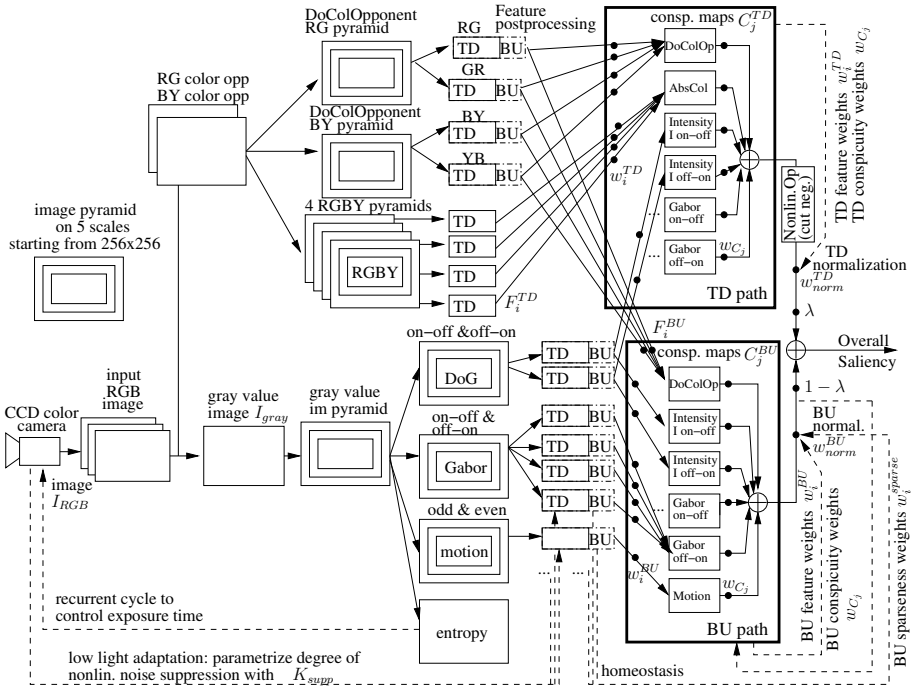


Fig. 1. Visual attention sub-system (dashed lines correspond to TD links).

supported (excitation), while the others are suppressed (inhibition) resulting in a sparse task-dependent scene representation. Opposed to that, the BU pathway (on bottom) supports an object- and task-unspecific filtering of input data supporting scene elements that differ from their surroundings. The BU pathway is important for a task-unspecific analysis of the scene supporting task-unrelated but salient scene elements.

The BU and TD saliency maps are linearly combined to an overall saliency on which FoAs are generated that determine the scene elements higher system layers will work on. The combination is realized using parameter λ (on the right hand side in Fig. 1) that is set dependent on the system state emphasizing the BU and/or TD influence. Due to this combination the system also detects scene elements that do not match the current TD system task and are hence suppressed in the TD pathway (to prevent inattentional blindness, i.e. complete perceptual suppression of scene elements as described in [12]).

Turning to the processing details, the following modalities are calculated on the captured color images: RGBY color (inspired from [7]), intensity by a Difference of Gaussian (DoG) kernel, oriented lines and edges by a Gabor kernel, motion by differential images and entropy using structure tensor.

In the following, these modalities are described in more detail, after which the entropy feature is specified that is used to set the camera exposure. The features motion and color are used differently for the BU and TD path. The BU path uses double color opponency from RGBY colors by applying a DoG on 5 scales on the RG and BY color opponent maps. The filter results are

separated into their positive and negative parts (on-off/off-on separation, whose importance is emphasized in [7]) leading to 4 pyramids of double color opponent RG,GR,BY and YB-channels. The TD path uses the same color feature but additionally 4 pyramids of the absolute RGBY maps. Absolute RGBY colors do not support the BU popout character and are hence not used in the BU path. A DoG filter bank is applied on 5 scales separating on-off and off-on effects. Furthermore a Gabor filter bank on 4 orientations (0, $\pi/4$, $\pi/2$, $3/4\pi$) and 5 scales is calculated separately for lines and edges (even and odd Gabor). The realized Gabor filter bank ensures disjoint decomposition of the input image. The detailed mathematical formulation of the used Gabor filter bank can be found in [13]. Motivated from DoG the concept of on-off/off-on separation is transferred to Gabor allowing e.g., the crisp separation of the sky edge or of street markings from shadows on the street. Motion from differential images on 5 scales is used in the BU path alone. Since this simple motion concept cannot separate static objects from self-moving objects, it is not helpful in TD search. The entropy T is based on the absolute gradient strength of the structure tensor A on the image I_{gray} (see Equation (1)). The matrix A is calculated using derivatives of Gaussian filters G_x and G_y and a rectangular filter of size W. We use the entropy as a means to adapt the camera exposure and not as a feature yet.

$$G_x(x,y) = -\frac{x}{2\pi\sigma^4} exp(-\frac{x^2+y^2}{2\sigma^2}), \quad G_y(x,y) = -\frac{y}{2\pi\sigma^4} exp(-\frac{x^2+y^2}{2\sigma^2})$$
$$A = \begin{bmatrix} \Sigma_W (G_x * I_{\text{gray}})^2 & \Sigma_W (G_x * I_{\text{gray}})(G_y * I_{\text{gray}}) \\ \Sigma_W (G_y * I_{\text{gray}})(G_x * I_{\text{gray}}) & \Sigma_W (G_y * I_{\text{gray}})^2 \end{bmatrix}, T = \frac{det(A)}{trace(A)} \quad (1)$$

The local exposure control works on the accumulated activation $T_{sum} = \Sigma_{RoI} T$ on an image region of interest (RoI) (e.g., coming from the appearance based object tracker that is part of our ADAS, for details see [8]). Here we get inspiration from the human local contrast normalization. The exposure time is recursively modified in search of a maximum on T_{sum}, which maximizes the contrast on the defined image regions. In sum, the system disposes of 130 independently weightable subfeature maps.

Following the calculation of the modalities a postprocessing step on all subfeature maps is performed (see Fig. 2). The feature postprocessing consists of 5

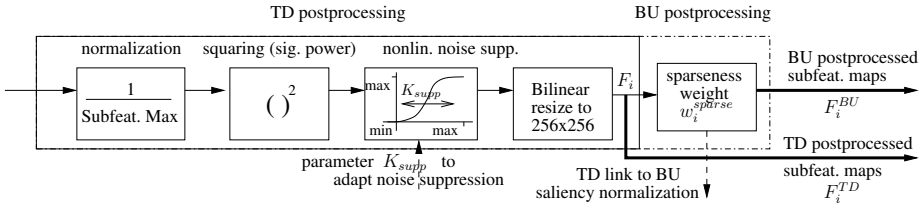


Fig. 2. Postprocessing of feature maps in BU and TD path.

steps. First all subfeatures are normalized to the maximal 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. The ideal input pattern is generated by setting all pixels to 1 whose matching pixel positions in the filter kernel are bigger than 0. Figure 3 shows the resulting ideal DoG and 0° even Gabor input patterns that are derived from the known filter kernels. This procedure ensures comparability between subfeatures of one modality. Next, the

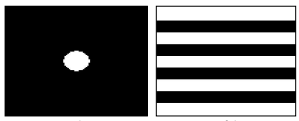


Fig. 3. Input patterns that maximize the filter response. The maximum of this filter response is used for subfeature normalization: a) Ideal DoG input pattern, b) Ideal 0° even Gabor input pattern.

signal power is calculated by squaring and a dynamic neuronal suppression using a sigmoid function is applied for noise suppression. A parameter K_{supp} shifts the sigmoid function horizontally, which influences the degree of noise suppression respectively the sparseness of the resulting subfeature maps. After a bilinear resize to the resolution 256x256 for later feature combination, for the BU feature postprocessing a sparseness weight w_i^{sparse} is multiplied that ensures popout by boosting subfeature maps with sparse activation (see Equation (2)).

$$w_i^{\text{sparse}} = \sqrt{\frac{2^s}{\sum\limits_{\forall x,y \text{ with } F_{i,k}(x,y)>\xi} F_{i,k}(x,y)}} \text{ for } s = [0,4] \text{ and } \xi = 0.9 \cdot Max(F_{i,k}) \tag{2}$$

The sparseness operator is not used in the TD path (see TD branch in Fig. 2) in order to prevent the suppression of weak object features.

Later in the TD path a weighting realizing inhibition and excitation on all 130 subfeature maps takes place. The TD weights w_i^{TD} are calculated in an offline step using Equation (3) (inspired by Frintrop [9]). The average activation in the object region is related to the average activation in the surround on each feature map F_i^{TD} taken only the N_i pixels above the threshold $K_{conj}Max(F_i^{TD})$ with $K_{conj} = (0, 1]$ into account:

$$w_i^{\text{TD}} = \begin{cases} \text{SNR}_i & \forall\ \text{SNR}_i \geq 1 \\ -\frac{1}{\text{SNR}_i} & \forall\ \text{SNR}_i < 1 \end{cases} \text{ with } \text{SNR}_i = \frac{\frac{\sum(F_{i,obj}^{\text{TD}} > K_{conj}Max(F_i^{\text{TD}}))}{N_{i,obj}}}{\frac{\sum(F_{i,surr}^{\text{TD}} > K_{conj}Max(F_i^{\text{TD}}))}{N_{i,surr}}} \tag{3}$$

In the BU path only excitation ($w_i^{\text{BU}} >= 0$) takes place, since without object or task knowledge in BU nothing can be inhibited. For a more detailed discussion of feature map weighting see [7, 8].

The subfeature normalization procedure ensures intra-feature comparability, but for the overall combination, comparability between modalities is required as well. We solve this 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 in biological systems (see e.g., [11]), 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 (4)), taking only the N_j pixel over the threshold $\xi = 0.9 \cdot Max(C_j^{\mathrm{BU}})$ into account. Exponential smoothing (see Equation (5)) 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.

$$\tilde{w}_{C_j}(t) = \frac{1}{\frac{1}{N_j} \sum\limits_{\forall x,y \text{ with } C_j^{\mathrm{BU}}(x,y) > \xi} C_j^{\mathrm{BU}}(x,y)} \text{ and } \xi = 0.9 \cdot Max(C_j^{\mathrm{BU}}) \tag{4}$$

$$w_{C_j}(t) = \alpha \tilde{w}_{C_j}(t) + (1-\alpha) w_{C_j}(t-1) \text{ for } j = 1..M \tag{5}$$

Before combining the BU and TD saliency using the parameter λ a final normalization step takes place. Like the subfeature maps, the saliency maps are normalized to the maximal expected value. For this we have to step back through the attention subsystem taking into account all weights ($w_i^{\mathrm{sparse}}, w_i^{\mathrm{BU}}, w_i^{\mathrm{TD}}, w_{C_j}$) and the internal disjointness/conjointness of the features to determine the highest value a single pixel can achieve in each conspicuity map. 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 disjoint between scales and orientations 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 product 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 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}. Using this approach w_{norm}^{TD} will adapt when the TD weight set changes.

3.2 Comparison to other TD attention models

Taken the abundance of computational attention models (see [14] for a review) we selected the two related approaches of Navalpakam [6] and Frintrop [7] for a

detailed structural comparison, since these impacted our work most. Then, we summarize what makes our approach particularly appropriate for the real world car domain.

The system of **Navalpakam** [6] is based on the BU attention model Neuromorphic Vision Toolkit (NVT) [10] but adds TD to the system. Each feature map is normalized to its current maximum, resulting in a loss of information about the absolute level of activity and a boosting of noise in case the activation is low. Taken such a normalization procedure and the object dependence of the TD weights, the BU and TD saliency maps are not comparable, since the relative influence of the TD map varies when the TD weight set is changed. Additionally, the BU and TD saliency maps are not weighted separately for combination. As features a speed-optimized RGBY (leading to an inferior separability performance), a DoG intensity feature and Gabor filter on 4 orientations (both without on-off/off-on or line/edge separation) are used on 6 scales starting at a resolution of 640x480. The system uses TD weights on all subfeature maps resulting in 42 weights that allow reasonable selectivity. A DoG-based normalization operator (see [10]) is applied for popout support and to diminish the noise resulting from the used feature normalization. However the absolute map activation is lost.

The system of **Frintrop** [7] integrates BU and TD attention and is real-time capable (see [15]). It was evaluated mainly on indoor scenes. The system normalizes the features to their current maximum, resulting in the same problems as described above. The BU and TD saliency maps are weighted separately for combination. Following the argumentation above the used normalization makes these combination weights dependent on the used TD weight set and thereby object-dependent. As features the system uses double color opponency based on an efficient RGBY color space implementation, a DoG intensity feature (with on-off/off-on separation), and a Gabor with 4 orientations starting from 300x300 resolution. A total of 13 TD-weights are used on feature (integrated over all scales) and conspicuity maps. For popout support a uniqueness operator is used.
Most important differences comparing the systems: We obtain high selectivity by decomposing the DoG (on-off/off-on separation) and Gabor (on-off/off-on separation, lines and edges) features without increasing the calculation time. Furthermore, the usage of TD weights on all subfeature maps and scales increases the selectivity. The resulting scale variance of the TD weights is not a crucial issue in the car domain. The RGBY is used as color and double color opponency. In contrast to [6, 7], we use motion to support scene dynamics. All subfeature maps and the BU respectively TD saliency maps are normalized without loosing information or boosting noise and by that preventing false-positive FoAs. Comparability of modalities is assured via homeostasis. The attention subsystem works on 5 scales starting at a resolution of 256x256. In the car domain bigger image sizes do not improve the attention system performance.

Our system supports conjunction of weak features since the sparseness operator is not used in the TD path. Illumination invariance is reached by image region specific exposure control that is coupled tightly to the system.

4 Results

In the following, we evaluate the system properties related to the challenges of Section 2. All results are calculated on five real world data sets (cars, reflexion poles, construction site, inner city stream, toys in an indoor scene) accessible in the internet (see [16]).

① **High feature selectivity:** In the car domain the search performance is strongly influenced by the horizon edge present in most images of highways and country roads. This serves as example problem for showing the importance of high feature selectivity. Typically, the horizon edge is removed by mapping out the sky in the input image, which might not be biological plausible. Based on the high selectivity of the attention features, we instead suppress the horizon edge directly in the saliency by weighting the subfeature maps. The gain of this approach is depicted in Fig. 4 that shows the diminished influence of the horizon edge on the (TD modified) BU saliency of the real world example in Fig. 5b). Table 1 shows the significant performance gain of attentional sky suppression versus no horizon edge handling on the average FoA hit number ($\overline{Hit}$) and detection rate ($\overline{DRate}$) (see [7] for definition of these measures) based on our real world benchmark data.

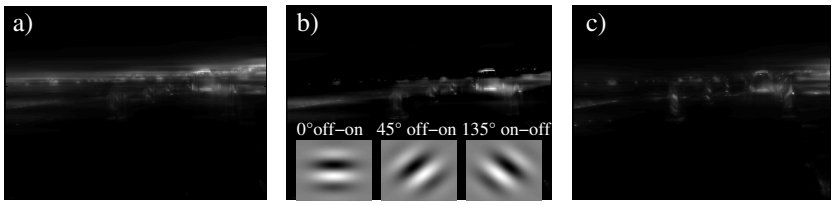


Fig. 4. Evaluation of selectivity (based on the input image depicted in Fig. 5b): a)Original BU saliency, b)modified BU saliency with attentional sky suppression (TD influence), using suppressive odd Gabor filter kernels in low scales, c)BU saliency, masked sky.

Search target	# test images	a) original BU $\overline{Hit}$ ($\overline{DRate}$)	b) attentional sky supp. $\overline{Hit}$ ($\overline{DRate}$)	c) sky masked $\overline{Hit}$ ($\overline{DRate}$)
Cars	54	3.06 (56.3%)	2.19 (71.4%)	2.47 (71.4%)

Table 1. Benefit of attentional sky suppression on real world data.

② **Comparable TD and BU saliency maps:** The used feature normalization prevents noise on the saliency map and ensures the preservation of the absolute level of feature activation. Using a TD weight set that supports certain 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 5f) shows that the maximal attention value on the TD saliency map for cars rises when the car comes into view (see [16] for downloadable result stream). The influence combining the now comparable TD and BU saliency maps is depicted in Tab. 2, showing that TD improves the search performance considerably. However, the influence of task-unspecific saliency (i.e., $\lambda < 1$) has to be preserved to avoid inattentional blindness.

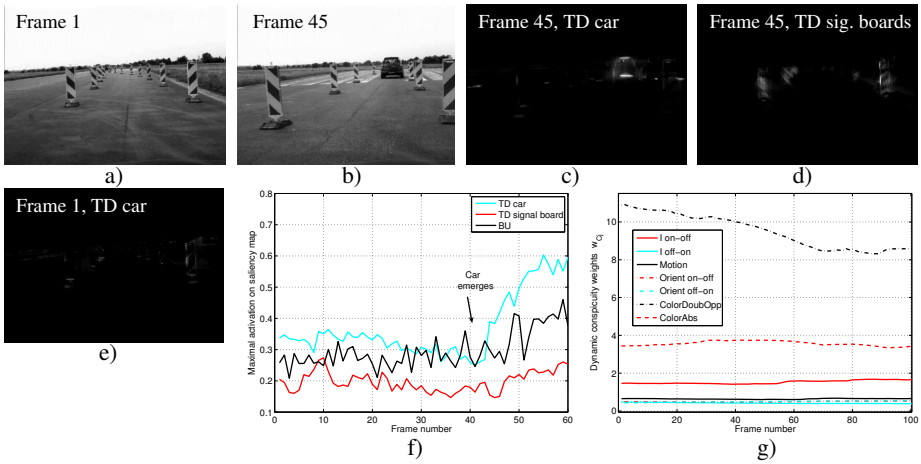


Fig. 5. Evaluation of normalization: a),b) Input images c)TD saliency tuned to cars, d)TD saliency tuned to signal boards, e)TD saliency tuned to cars (noise, since no car is present), f)Maximal saliency activation level on BU, TD car and TD signal board map, g) Dynamically adapted conspicuity weights w_{C_j} (homeostasis) for the M=7 modalities.

Target	# Test im (obj)	# Training im	Aver. FoA hit number (and detection rate [%]) $\lambda = 0$ (BU)	$\lambda = 0.5$ (BU & TD)	$\lambda = 1$ (TD)
Cars	54 (58)	54 (selftest)	3.06 (56.9%)	1.56 (93.1%)	1.53 (100%)
		3	3.06 (56.9%)	1.87 (89.7%)	1.82 (96.6%)
Reflection poles	56 (113)	56 (self test)	2.97 (33.6%)	1.78 (59.8%)	1.85 (66.3%)
		3	2.97 (33.6%)	2.10 (51.3%)	2.25 (52.2%)

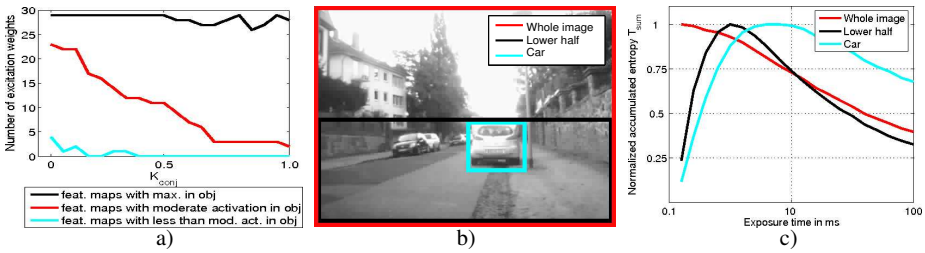
Table 2. Linear combination of BU and TD saliency, influence on search performance ($\lambda = 0$ equals pure BU and $\lambda = 1$ pure TD search)

③ Comparability of modalities: The used dynamic adaptation of w_{C_j} (homeostasis, see Equation (5)) causes a twofold performance gain. First, the a priori incomparable modalities can be combined 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 R color channel will not be overrepresented in the saliency). Figure 5g) depicts the dynamically adapted w_{C_j}. Table 3 shows a noticeable SNR gain on the overall saliency for 26 traffic relevant objects (e.g., traffic light, road signs, cars), comparing the dynamically adapted w_{C_j} vector with a locally optimized static w_{C_j} vector.

Traffic-relevant objects	#images (obj)	$\overline{\text{SNR}}_{obj}$ 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 3. Comparability of modalities via homeostasis.

④ Support of conjunctions of weak object features in the TD path is assured since w_i^{sparse} is used in BU only. Evaluation on 54 images with cars as TD search object shows that the average object signal to noise ratio ($\overline{\text{SNR}}_{obj}$)



a) b) c)

Fig. 6. Evaluation of illumination influence: a)# of excitatory TD weights depending on K_{conj}, b)Image regions used for exposure optimization (whole image, lower half, car), c)Energy function: Accumulated entropy T_{sum} with object-dependent optima.

on the TD saliency map (defined as the mean activation in the object versus its surround) decreases by 9% when w_i^{sparse} is also used in the TD path. For evaluation we define weak object feature maps as having the current maximum outside the object region but still having object values of at least 60% of the maximum within the object. For the used 54 traffic scene images 11% of all feature maps are weak. In case weak feature maps are used to optimally support the TD saliency in an excitatory way $\overline{\text{SNR}}_{obj}$ on the TD saliency map increases by 25%. The results are aggregated in Tab. 4. Figure 6a) shows that the number of excitatory TD weights w_i^{TD} decreases the bigger K_{conj} is. An object-dependent trade-off exists since the TD saliency map gets sparser the bigger K_{conj} is.

TD search target	# test image	$\overline{\text{SNR}}_{obj}$ with w_i^{sparse}	$\overline{\text{SNR}}_{obj}$ without w_i^{sparse}	$\overline{\text{SNR}}_{obj}$ with optimal weak feat. excitation
Cars	54	6.72	7.32 (+9%)	8.41 (+25%)

Table 4. Improvement due to support of weak feature conjunctions.

⑤ **Changing lighting conditions:** The feature activation of an image region depends on the illumination. Hence the TD weight set is only optimal for the lighting conditions of the training images and the TD search performance decreases when illumination changes without an adaptation of the camera exposure. It is important to note that in a real world scene the optimal exposure in varying illumination is different for all objects (see Fig. 6b and c), making the exposure control dependent on the current task of the system. Evaluation based on a complex indoor test setting where we were able to control the illumination shows that the realized exposure control leads to illumination invariance of the TD weight sets (see Tab. 5).

Target	# Test	Average hit number (and detection rate [%]), TD search $\lambda = 1$				
Toys in a	im (obj)	Traning illu-	without expos. control		with expos. control	
complex in-		mination 75 lx	150 lx	15 lx	150 lx	15 lx
door setup	20 (20)	1.95 (100%)	2.74 (95%)	2.83 (30%)	1.80 (100%)	2.0 (100%)

Table 5. Illumination invariance of TD weight sets using dedicated exposure control.

5 Summary

This paper describes a flexible biologically motivated attention subsystem that is used as the front end of an ADAS. The real world requirements of the car domain

have resulted in an improved system performance by incorporating modulating TD links.

The key enhancements of our attention subsystem are: high feature selectivity, a normalization leading to comparable BU and TD saliency maps enabling their combination with a linear combination weight λ that is independent of the used TD weight set, the comparability of the used modalities, the conjunction support of weak object features in the TD path and an exposure control that depends on the object in focus or a task relevant image region. These principles lead to a robust system suitable for the dynamic real world environment.

Using a purely vision based situation analysis that worked on TD prefiltered vision data provided by an earlier version of the attention subsystem described here, we showed the real time capability of our system in a real world test setup, where our prototype car was reliably able to brake autonomously (see [8]).

References

1. Corbetta, M., Shulman, G.: Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews Neuroscience **3** (2002) 201–215
2. Egeth, H.E., Yantis, S.: Visual attention: control, representation, and time course. Annual Review of Psychology **48** (1997) 269–297
3. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nat. Reviews Neuroscience **5**(6) (June 2004) 495–501
4. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology **4**(4) (1985) 219–227
5. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial Intelligence **78**(1-2) (1995) 507–545
6. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Research **45**(2) (2005) 205–231
7. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. PhD thesis, University of Bonn Germany (2006)
8. Michalke, T., Gepperth, A., Schneider, M., Fritsch, J., Goerick, C.: Towards a human-like vision system for resource-constrained intelligent cars. In: Int. Conf. on Computer Vision Systems, Bielefeld (2007)
9. Frintrop, S., Backer, G., Rome, E.: Goal-directed search with a top-down modulated computational attention system. In: DAGM-Symposium. (2005) 117–124
10. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11) (1998) 1254–1259
11. Hardy, R.N.: Homeostasis. Arnold (1983)
12. Simons, D., Chabris, C.: Gorillas in our midst: Sustained inattentional blindness for dynamic events. British Journal of Developmental Psychology **13** (1995) 113–142
13. Trapp, R.: Stereoskopische Korrespondenzbestimmung mit impliziter Detektion von Okklusionen. PhD thesis, University of Paderborn Germany (1998)
14. Heinke, D., Humphreys, G.: Computational models of visual selective attention: a review. In Houghton, G., ed.: Connectionist Models in Psychology, Psychology Press (2005) 273–312
15. Frintrop, S., Klodt, M., Rome, E.: A real-time visual attention system using integral images. In: Int. Conf. on Computer Vision Systems, Bielefeld (2007)
16. BenchmarkData: (2007) `http://www.rtr.tu-darmstadt.de/~tmichalk/ICVS2008_BenchmarkData/`.