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

Task-Based Environment Interpretation and System Architecture for Next Generation ADAS

Robert Kastner, Thomas Michalke, Jürgen Adamy, Jannik Fritsch, Christian Goerick

2011

Preprint:

This is an accepted article published in IEEE Intelligent Transportation Systems Magazine,. The final authenticated version is available online at: https://doi.org/[DOI not available]

Task-Based Environment Interpretation and System Architecture for Next Generation ADAS

©PHOTODISC

Digital Object Identifier 10.1109/MITS.2011.942201
Date of publication: 28 October 2011

Robert Kastner
Honda R&D Europe (Deutschland) GmbH
robert kastner@de.hrdeu.com

Thomas Michalke
Darmstadt University of Technology
thomas paul.michalke@daimler.com

Jürgen Adamy
Darmstadt University of TechnologyInstitute for Automatic Control
adamy@rtrr.tu-darmstadt.de

Jannik Fritsch and Christian Goerick
Honda Research Institute Europe GmbH, Germany
{jannik.fritsch,christian.goerick}@honda-ri.de

1939-1390/11/$26.00©2011IEEE

I. Introduction

From our point of view current research in the area of Advanced Driver Assistance Systems (ADAS) is focused mostly on single, independent, highly specialized tasks. To this end, today's Driver Assistance Systems are engineered for supporting the driver in clearly defined traffic situations like, e.g. keeping a specified distance to the vehicle in front. Some may argue that the quality of an engineered system in terms of isolated aspects (e.g. object detection or tracking) is often sound, but the solutions lack the 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.

***Abstract*–**State-of-the-art advanced driver assistance systems (ADAS) typically focus on single tasks and therefore, have clearly defined functionalities. Although said ADAS functions (e.g. lane departure warning) show good performance, they lack the general ability to extract spatial relations of the environment. These spatial relations are required for scene analysis on a higher layer of abstraction, providing a new quality of scene understanding, e.g. for inner-city crash prevention when trying to detect a *Stop* sign violation in a complex situation. Otherwise, it will be difficult for an ADAS to deal with complex scenes and situations in a generic way. This contribution presents a novel approach of task-dependent generation of spatial representations, allowing task-specific extraction of knowledge from the environment based on our biologically motivated ADAS. The approach also incorporates stored knowledge in form of digital map data, introducing a new way of eHorizon integration. Additionally, the hierarchy of the approach provides advantages when dealing with heterogeneous processing modules, a large number of tasks and additional new input cues. The results show the reliability of the approach and also the increase of performance on the system level.

***Index Terms*–**driver assistance, scene analysis, environment representation.

Additionally, the combination of numerous dedicated algorithms for virtually all existing tasks/objects/classes (each focusing on a single aspect of an ADAS) is not feasible in terms of processing power. From our point of view, a generic vision-based scene decomposition is necessary to cope with the limited amount of computational resources. Vision systems inspired by biology turned out to be highly flexible and also capable of adapting to severe changes in the task and/or the environment. One of our design goals on our way to achieve such an "all-situation" ADAS is to implement a biologically motivated, cognitive vision system. This vision system is the perceptual front-end of an ADAS, which can handle the wide variety of situations typically encountered when driving a car. For more information on this kind of vision system refer to [1].

Another important issue of system design is the proactive nature of a system. In this context proactive means: the capability of a system to actively decide based on the current system state and sensor input, which task to handle next. Otherwise, it will be challenging to deal with all tasks at the same time. Therefore, an in-depth understanding of the current scene is necessary making scene analysis even more relevant. Details about the proactive extension of our biologically motivated system design can be found in [2].

The main intention of this contribution is to present a generic way for representing and combining extracted spatial knowledge of the environment. Spatial knowledge describes the relation between position, size, type and movement of objects and road (e.g. relating a vehicle to a lane). Additionally, the incorporation of stored knowledge in form of digital map data is also introduced. Recently, in the scientific community the field of designing and researching spatial representations has gained interest. In most of the related research some kind of probabilistic grid is used to integrate information from sensors over time. Hence, spatial information of occupied areas within the surrounding can be provided (see [3] for one of the early approaches). Also, numerous contributions have shown the extraction of moving objects, cars, etc. from an occupancy grid (see e.g. [4]). Nevertheless, in most cases the spatial representation is only capable of storing and interpreting the low level information of some kind of sensor like for example a laser scanner. Therefore, it is difficult to easily integrate results of other algorithms (like traffic sign recognition) in a generic way. As opposed to that, our aim is to provide a generic method for the combination of different processing results as well as stored knowledge, exploiting spatial relations on a higher level of abstraction.

To put it differently, this contribution focuses on a generic way to combine the results of different processing modules in order to extract task-specific knowledge of the environment based on spatial representations. The goal is to develop a cognitive system that is able to combine spatial knowledge of the environment depending on the current task. The idea of using spatial representations was inspired by C. Colby [5], who showed that the human brain constructs multiple spatial representations, because each eases a certain task. To our knowledge, there is no automotive approach that is able to integrate different types of processing results in a generic

Spatial relations allow for scene analysis on a higher layer of abstraction, providing a new quality of scene understanding.

way. With such an approach the extraction of information on a higher level of abstraction becomes possible. The realized system is able to deal with complex scenes and generate spatial expectations for the current task. The system is tested on real-world data and qualitative results as well as quantitative performance improvements are shown. To this end, the performance increase of single algorithms if integrated in the system is shown by using generic temporal integration and fusion procedures.

II. Related Work

The topic of researching intelligent cars is gaining interest as documented by the DARPA Urban Challenge [6] and the European Information Society 2010 *Intelligent Car Initiative* [7] as well as several European Projects like, e.g., Safespot or PReVENT.

The INSAFES project treats the integration of safety applications within the project PReVENT. Hence, Amditis et al. [8] provide an approach for the integration of ADAS functions, by using a single perception layer as well as a common action layer. The function layer in between the two layers is different, also allowing parallel applications that do not have a unified reasoning, therefore weakening the idea of integration.

Publications that deal with spatial representations, in general, are quite numerous. Nevertheless, a lot of these contributions use an evidence grid to integrate sensor data over time (see e.g. [9]). An evidence grid provides a framework for a probability-based approach, where the occupancy of a cell is transformed to a likelihood. Therefore, the main task is to provide the free driving space. Other approaches focus on the fusion of two evidence grids as for example [10]. Additionally, the authors propose an efficient map data structure called Deferred Reference Count Octree (DRCO), solving storage problems when using 3D evidence grids. Also common is the extraction of knowledge from an evidence grid, as e.g. done by [11], proposing a method for distinguishing between static and dynamic objects when building an environment map. To this end, the focus of publications regarding probabilistic grids is mainly on sensor fusion, temporal integration and knowledge extraction from sensor data. In contrast, our work allows a combination of results from different heterogeneous processing modules at higher processing levels. More specifically, we extract spatial relations from the combination of different processing results, instead of directly interpreting sensor data as done by other approaches. Nevertheless, the free area from an evidence grid can also be used as an input result for our task-dependent representation generation.

Several publications deal with the modeling of the environment in a graph-based manner as for example [12]. But mainly with the intention of a compact representation for a digital map database, instead of extracting spatial relations from processing results in a generic way.

As stated before, a novel approach of spatial representation is introduced, which is embedded into a biologically inspired ADAS. Turning to biological vision system as a part of this rather new domain of research multiple mildly related approaches exist. One of the most prominent examples is a system developed in the group of E. Dickmanns [13]. It uses several active cameras mimicking the active nature of gaze control in the human visual system. But no tuneable attention system and no top-down aspects are incorporated as existing in the human visual system.

An artificial vision system in the vehicle domain that also includes an attention system and that hence is somewhat related to our approach is described in [14]. The approach allows for a simple bottom-up attention-based decomposition of road scenes but without incorporating object or prior knowledge. Therefore, the system is not able of an in-depth scene analysis using spatial relations as the here proposed system. For a more detailed comparison to the state-of-the-art in human-like vision systems, refer to [15].

To our knowledge, in the car domain no biologically motivated large scale systems exists that allows task-dependent evaluation based on spatial representations.

III. System Description

The proposed overall architecture concept for a biologically motivated system design incorporating task-dependent scene analysis is depicted in Fig. 1. It consists of four major parts: the “static domain-specific tasks”, the “what” pathway, the “where” pathway, and a part allowing “environmental interaction”.

The distinction between “what” and “where” processing path is motivated from the human visual system where the dorsal and ventral pathway are typically associated with these two functions (see, e.g. [16]). 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 [16]). Nevertheless, an ADAS also requires context information in the form of the road, its shape, and the current global scene context (e.g. inner-city), generated by

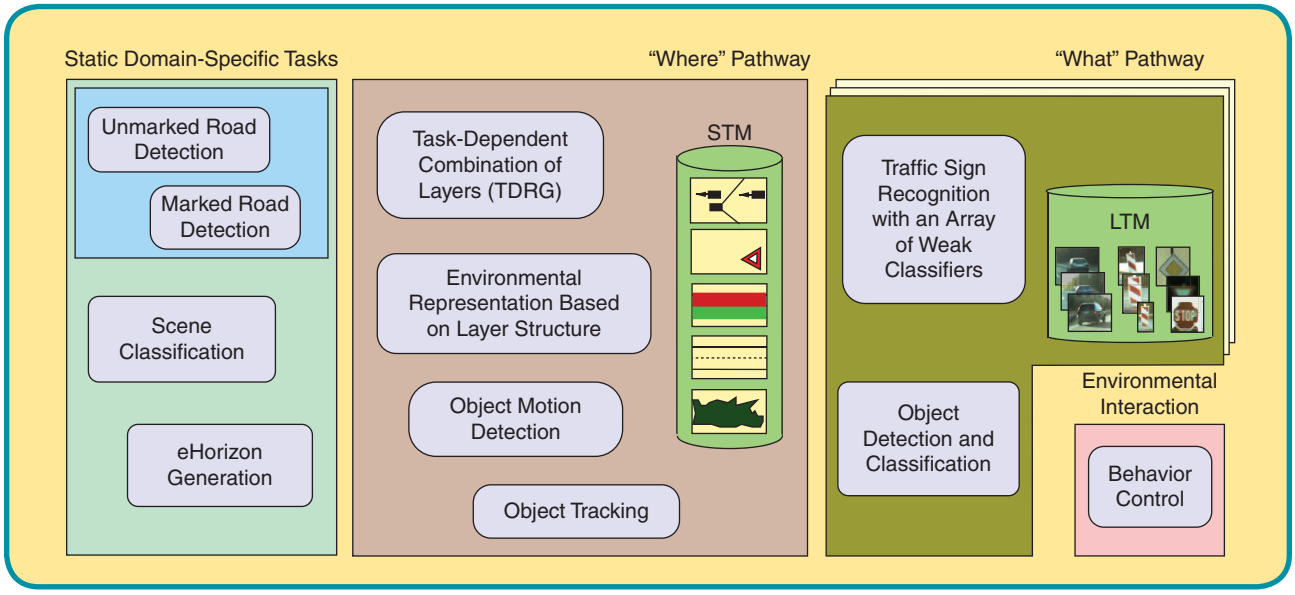


FIG 1 Overall system architecture for task-dependent scene analysis.

the static domain-specific part. Furthermore, for assisting the driver, the system requires interfaces for allowing environmental interaction (i.e. triggering actuators).

In order to allow an understanding of the proposed task-dependent representation generation a rough system description is given (for more details on these system modules refer to [15] and the following references in the subsections). In Section III-D, the task-dependent representation generation is explained in detail.

A. The "What" Pathway

1. Detection

Starting in the "what" pathway (see Fig. 2) the 400×300 pixel color input image (Fig. 3a) is analyzed by calculating the so-called saliency map S^{total} (see Fig. 3b and c for a visualization of S^{total}). The saliency map allows the complexity-minimizing decomposition of the visual scene. It results from the so-called attention principle. The attention is a generic information preprocessing principle that is believed to exist in the visual pathway of the mammal brain. The involved brain areas allow the prefiltering of the sensed environment (e.g., visual, acoustic, olfactory channel) in order to minimize its complexity. It is believed that with (1) top-down respectively (2) bottom-up driven attention two separate preprocessing principles exist. These principles allow a (1) task-specific respectively (2) purely sensory, task-unspecific attention generation.

The saliency map S^{total} combines the task-specific and unspecific attention generation and results from a weighted linear combination of $N = 130$ biologically inspired input feature maps F_i. 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 [17] as attention feature that models the processing of photoreceptors on the retina.

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. The weights w_i^{TD} dynamically boost feature maps that are important for our current task or object class in focus and suppress the rest. The bottom-up (BU) weights w_i^{BU} are set object-unspecifically in order to detect unexpected potentially dangerous scene elements. The parameter $\lambda \in [0, 1]$ determines the relative importance of TD and BU search in the current system state. For more details on the attention system please refer to [18].

2. Classification

Now, we compute the maximum on the current saliency map S^{total} and get the focus of attention (FoA, i.e., the currently most interesting image region) by generic region-growing-based segmentation on S^{total}. In the following, with the FoA a restricted part of the image is classified using a state-of-the-art object classifier that is based on neural nets [19]. Different from the generic classifier concept present in the "what" pathway, traffic signs are treated separately with an array of weak classifiers for classification as described in [20]. The array of weak classifiers is similar to the idea of Viola and Jones [21]. Hence, a probability value for each of the sign classes is computed for all provided FoAs which were generated by the attention system. Therefore, j independent weak classifiers compute a probability P_j^{FoA} that indicates the existence of a certain traffic-sign-class-specific attribute at

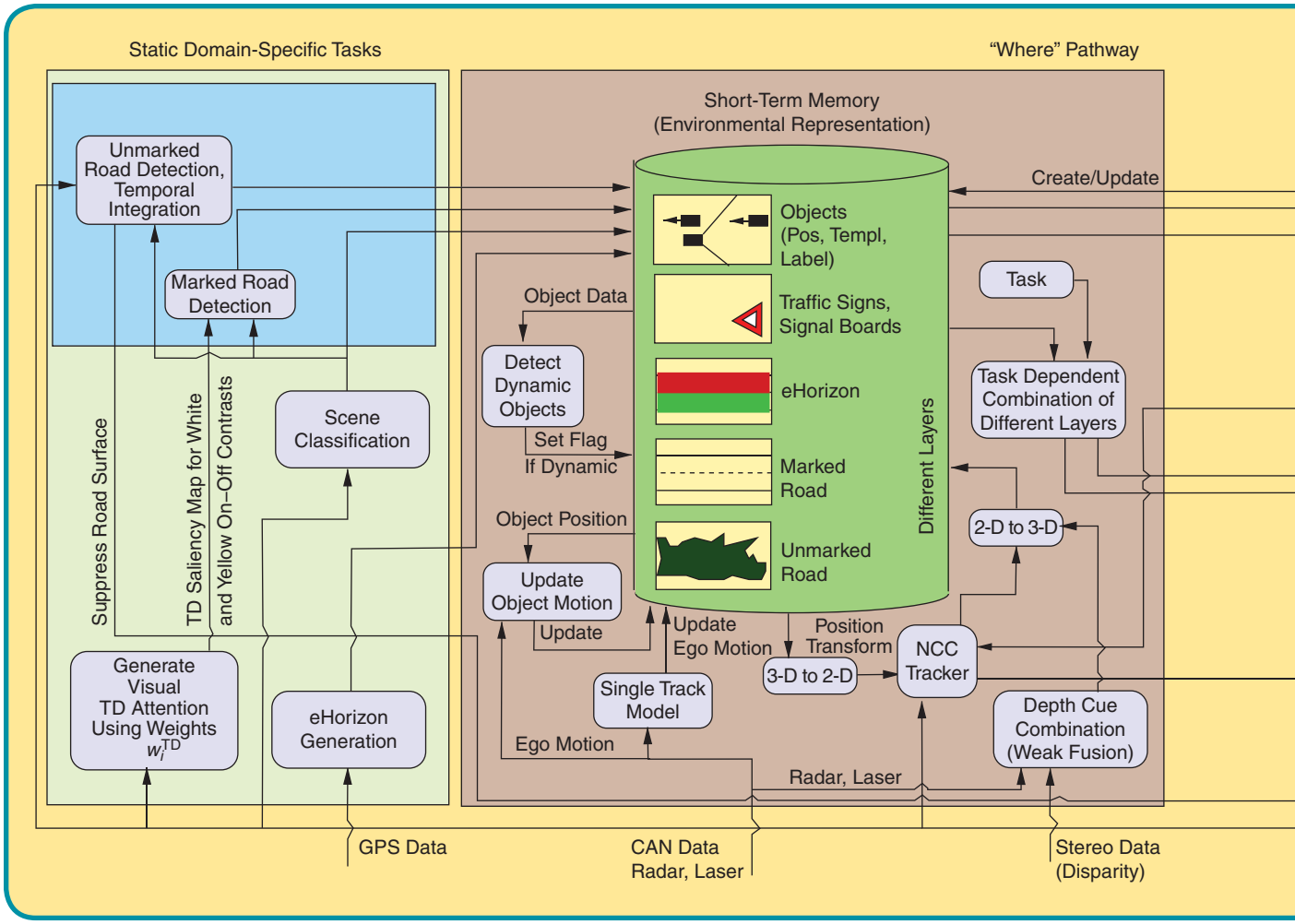


FIG 2 Biologically motivated system structure for task-dependent scene analysis using spatial representations and stored knowledge.

each FoA. The weak classifiers are based on the following attributes of the traffic sign classes: color, corner matching, height in the world, pixel relation, excentricity, corner relation and shape (see [20] for details).

The overall 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.

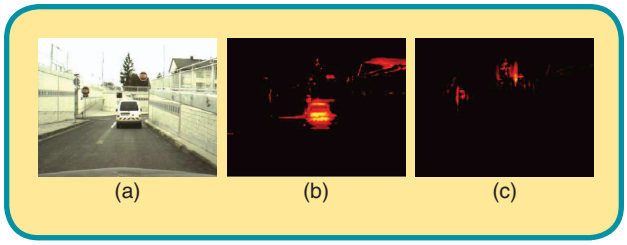
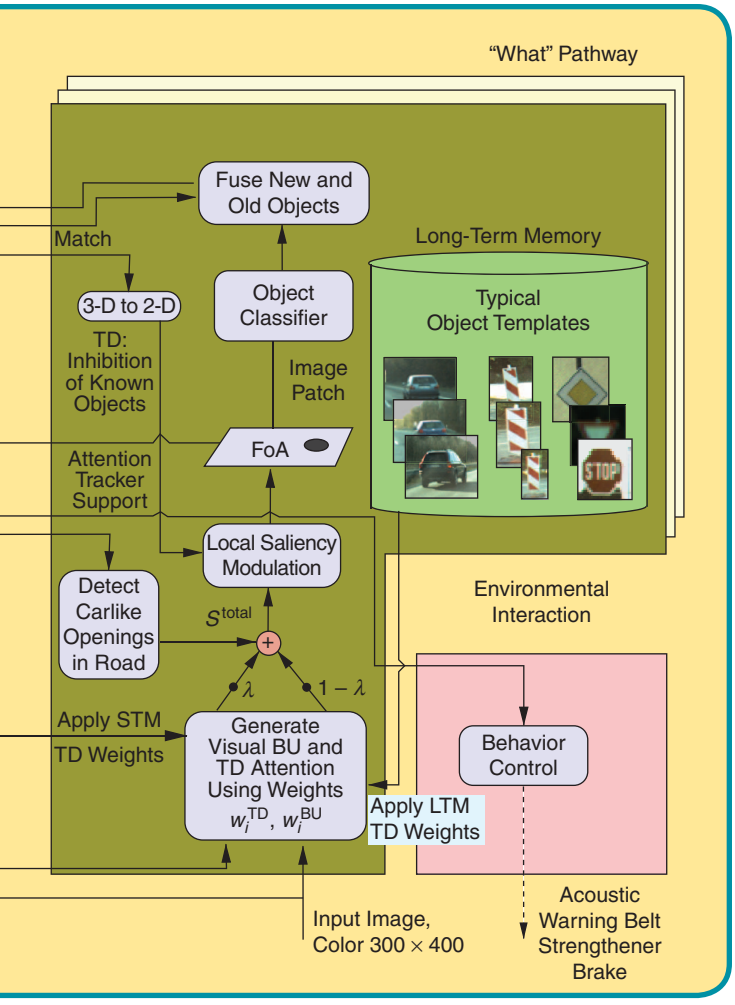


FIG 3 Attention-based scene decomposition: (a) Inner-city scene, (b) TD attention tuned to cars (S^{total}), (c) TD attention tuned to traffic signs (S^{total}).

3. Information Fusion

Internal information fusion processes improve the performance of system modules. For example, the certainty for a recognized traffic sign can be improved by a fusion with the digital map data. The digital map data is also called electronic horizon (eHorizon), since it provides a cut-out of the complete digital map data based on the current position and driving direction. Therefore, the eHorizon provides information of static objects and road in the current environment. The result of the array of weak classifiers can be compared with the eHorizon (see Section III-B) to boost the confidence. If the type of traffic sign matches between recognition and map data the confidence is increased with a non-linear distance metric. Furthermore, the detected road (see Section III-B) is fused as context information into the attention system. More specifically, the road is suppressed in all feature maps F_i before fusing them in the overall saliency S^{total}. This procedure makes the saliency map S^{total} sparse and improves the TD weight quality. Additionally, TD-links are used for the modulation of the attention based on detected car-like openings in the found



drivable road segment. These car-like openings are detected by searching for car-sized openings in the road segment (see [15] for details). Additionally, the task-dependent layer combination can further focus the searched road area to e.g. the ego-lane (see Section III-D). Also, the information of the digital map data is used to actively search for specific objects in the current scene as for example traffic signs and road markings.

4. Long Term Memory

Finally, the "what" pathway contains a long term memory (LTM) that stores the generic properties of object classes. The LTM is filled offline with typical patches and corresponding aggregated feature map activations for all supported object classes. Based on that data, an online computation of the TD attention weights w_i^{TD} gets possible, thereby allowing the active search for virtually all possible object classes. Currently, we use cars, signal boards and a number of traffic signs as LTM content, although our system is not restricted to these object classes (see [18]). It is important to note that multiple LTM object classes are searched at the same time, which requires several "what" pathways running in parallel (depicted on Fig. 2 as multiple "what" pathways). In the default case, a specific "what" pathway searches for one LTM object type. This is done by computing the geometric mean of all TD weight sets of the specific LTM objects.

B. Static Domain-Specific Tasks

In the following part, the domain-specific tasks are described. These are the marked and unmarked lane detection, a reliable scene classification and a digital map provider (also called eHorizon).

1. Marked Lane Detection

The marked lane detection is based on a standard Hough transform whose input signal is generated by our generic attention system. The TD attention weights used here boost white and yellow structures on a darker background (so called on-off contrast), to which the biological motivated DoG filter is selective. The yellow on-off structures are weighted stronger than the white to allow the handling of lane markings in construction sites. The filtered result of the TD attention is transformed to the bird's eye view (i.e., the view from above, refer to [22] for details) before applying the Hough transform. Therefore, a clothoid model-based approach for detecting the markings is used (see, e.g., [23], [24], [25] for related clothoid based approaches). But with the knowledge of the current scene context and eHorizon (see later in this sub-section) a prior for the scene-specific lane width/position is set for the evaluation of the Hough space (e.g. when using the scene context a lane width of around 3.7 m is expected for highways). To this end, the result of the marked lane detection is related to metric coordinates directly suitable for a task-dependent representation.

2. Unmarked Lane Detection

The state-of-the-art unmarked lane 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 in the street training region (stereo, edge density, color hue, color saturation) 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. A temporal integration procedure between the current and past detected road segments based on the bird's eye view is applied. The procedure is used to increase the completeness of the detected road by decreasing the number of false negative road pixels (refer to [26] for a comprehensive description of the overall procedure). The result of the unmarked road detection is also in metric coordinates.

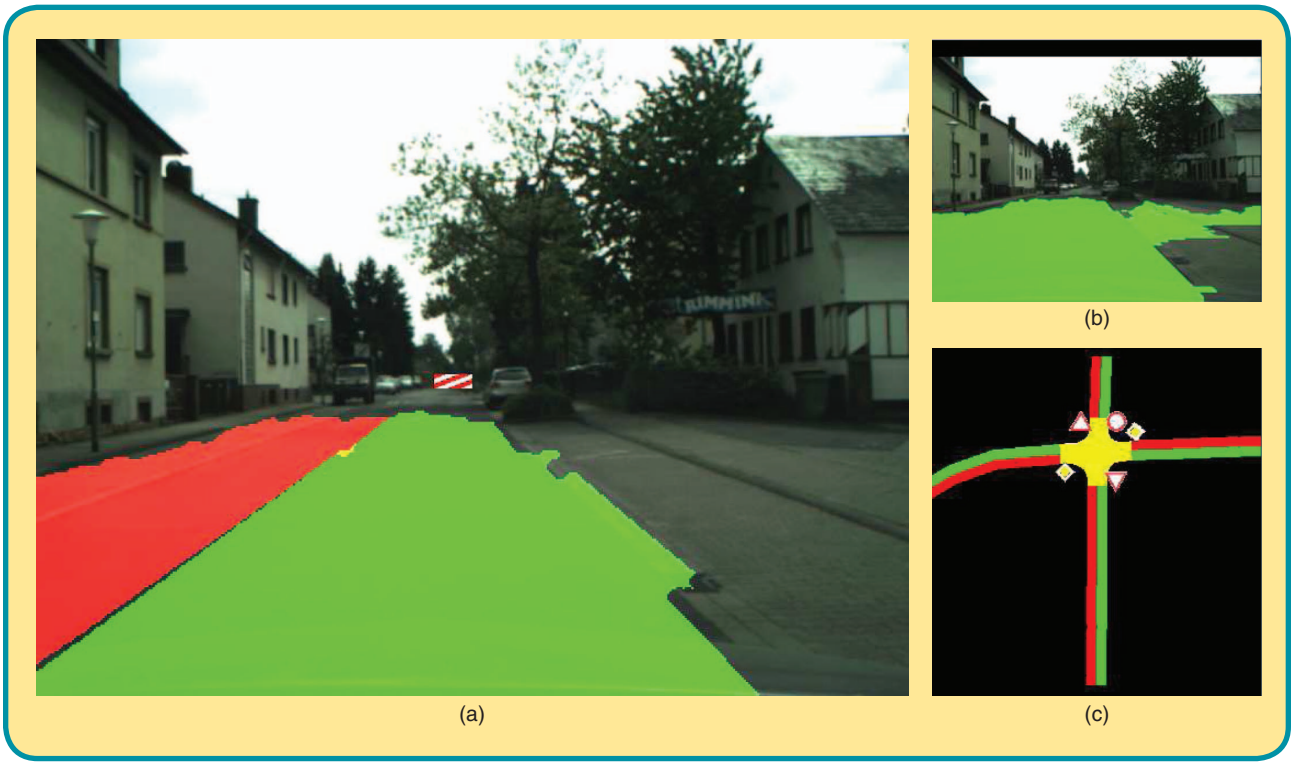


FIG 4 First qualitative evaluation example: (a) Image with results of all tasks based on the unmarked road detection result and data of the eHorizon. (b) Image with the unmarked road detection result in green. (c) Visualization for the spatial map provided by the eHorizon.

3. Digital Map

Furthermore, a short introduction to the used digital map provider (eHorizon) and the data it is supplying is given. Different from the related work (as, e.g. [27]), our eHorizon does not provide map data in a preprocessed way (e.g. length of a road segment with curvature) as usally done when directly focusing on a specific functionality. To this end, our digital map database does not incorporate environment models to reduce the amount of data. Hence, our eHorizon provides a metric map with a resolution of 0.1 m, in form of a local environmental map (see Figure 4c for a visualization). This has the advantage that no constraints regarding the road topology have to be made, which could lead to problems in inner-city areas. Additionally, the data of the eHorizon can be directly used as a virtual sensor. Therefore, the digital map (eHorizon) is not limited to a single functionality, but provides the basis for a number of tasks when used with the task-dependent representation generation.

The digital map database is generated by manual annotation of high resolution aerial images. Regarding the availability of the map data, one could also think of an automated way for the annotation by use of specific algorithms, as recently shown by Pink and Stiller [28].

Similarly to other eHorizon providers a precise GPS position, odometry data and map matching is used to generate a position and heading for the extraction of the information from the database (see e.g. [29]). To this end, the information supplied by the eHorizon is a spatial map, which shows the environment ahead of the driving direction.

The information which can be provided by the digital map is the following: geometry and direction of lanes (number), intersections, traffic signs and traffic lights. Nevertheless, also addtional information can be easily stored/encoded in the database.

4. Visual Scene Classification

The final part of the static domain-specific tasks is the state-of-the-art visual scene classification. For being able to run different modes of operation the current scene context (e.g. inner-city, country road, highway) has to be known. Otherwise it is not possible to parameterize the processing modules, as well as the task-dependent representation generation to the global characteristics and driving rules of the scene. For the computation of the scene classification only a single image is required as input, the processing is roughly the following: After the preprocessing the resulting image is divided in 16 parts and each part is independently transformed to the frequency domain. In the following, each transformed part is sampled with an array of shifted and oriented Gaussian filters, resulting in an average power spectrum for each of the parts. Finally, the classification is done with the Hierarchical Principal Component Classification, having learned during a

training phase a classification tree structure, based on the average power spectra of all parts. For more information please refer to [30].

C. Environmental Interaction

The system can interact with the world via an actuator control module. For example, for an emergency braking depending on the distance and relative speed of a recognized obstacle, the system can use a three phase danger handling scheme as shown in earlier versions of the here described ADAS (see [1]).

D. The "Where" Pathway

The central element for the task-dependent representation generation is the "where" pathway, providing on the one hand the basic spatial representations by the short term memory (STM) with generic update, fusion and temporal integration procedures. And on the other hand, the task-dependent combination of different representation layers. First the structure of the STM and its procedures will be described and afterwards the concept of task-dependent combination of different representation layers.

1. Short Term Memory

Starting with the former, the STM contains different layers, which are used to store different classes (see Fig. 2, STM within the "where" pathway sub-graph). The update/fusion process is strongly simplified due to the fact that only elements of the same class have to be treated. Furthermore, each layer has the same size and is a metric representation of the current environment (for one particular class) as seen from above. Therefore, the height of elements will not be depicted, but the different class layers roughly reflect different height levels of the world. The hierarchical order of the classes is the following, starting with the unmarked road layer as the lowest layer and finally, the highest layer is the object layer. At each time step (on the basis of the image recording frequency) all elements (on each layer) will be shifted and rotated according to the ego movement of the car, based on a Kalman filter prediction.

2. Object Fusion

The next step is the fusion between a newly detected object O_{new} and the already known ones. Depending on the class of the newly detected object either the traffic sign layer or the object layer is chosen. Based on the 3D position and size of O_{new}, a radius in the corresponding class layer of the STM is searched. If there is no other object within the radius, the layer is updated with the newly detected object. Otherwise, the object O_f found within the radius is compared to the new object O_{new} by means of the distance measure $\delta(O_f, O_{\text{new}})$ 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 $H_i^{O_{\text{new}}}$ (see Eq. (1)).

$$\delta(O_f, O_{\text{new}}) = \sum_{i=1}^{N} \sqrt{1 - \gamma(H_i^{O_f}, H_i^{O_{\text{new}}})}$$
$$\gamma(H_i^{O_f}, H_i^{O_{\text{new}}}) = \sum_{\forall x,y} \sqrt{H_i^{O_f}(x,y) H_i^{O_{\text{new}}}(x,y)}. \quad (1)$$

If the similarity exceeds a certain class-specific threshold, the new position will be stored in the associated layer of the short term memory (STM). Despite the initial validity of an object (after a first recognition) also a temporal integration scheme can be used. If activated an object has to be found a predefined number of times in consecutive frames before it is valid, reducing false positive detections. The valid objects in the STM are then suppressed in the current calculated saliency map to enable the system to focus on new objects. The principle of suppressing already detected and hence known objects was proven to exist in the human vision system and is termed inhibition of return (IoR), refer to [31] for details.

3. Object Tracking

All valid known objects and traffic signs are tracked using a 2D tracker that is based on normalized cross correlation. The tracker gets its anchor (i.e. the 2D pixel position where the correlation-based search for an object will be started in the new image) from a Kalman-filter-based prediction on the 3D representation taking the ego-motion of the camera vehicle and tracked object into account. This is a generic process and therefore, can be applied to any newly added class layer (see [2] for details).

A comparison between the current Kalman-fused 3D object position and the predicted object position (derived from the measured vehicle ego motion) allows the classification of detected objects as static/dynamic (see [15] for details).

If 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 interrupts the standard processing in the specific "what" pathway and 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 O_s. The object O_f found by the STM search is then compared to the searched object O_s again by means of the distance measure $\delta(O_f, O_s)$ based on the Bhattacharya coefficient as already described (see Eq. (1)).

4. Task Dependent Representation Generation

In the following, the concept of task-dependent combination of different representation layers is explained. Therefore, Fig. 5 shows the strongly simplified system structure with the used hierarchy for the layer combination. The system structure of Fig. 2 is visually simplified to five processing modules providing the input for the STM on Fig. 5. Nevertheless, the functionality remains as already described. Therefore, the subsequent task-dependent combination of layers is shown in more detail.

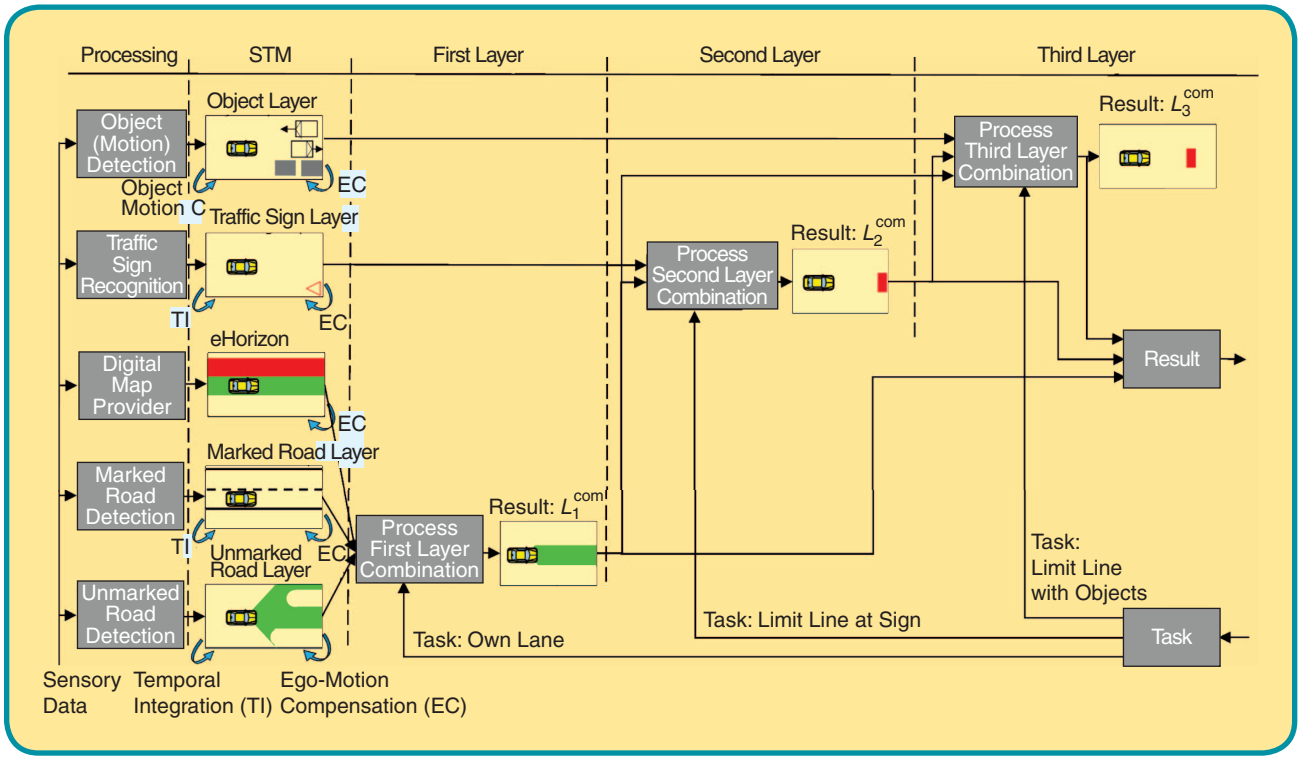


FIG 5 Concept of task dependent representation generation.

The unmarked road layer (L_{UR}), marked road layer (L_{MR}) and eHorizon layer (L_{eH}) are always combined as 1st layer combination (L_1^{com}, see Fig. 5, 1st layer), whereas the following combinations (2nd/3rd layer) only depend on the current task. The eHorizon provides detailed information on the area as well as driving direction of the lanes, therefore the desired lane(s) can be simply queried from the L_{eH} layer. Since, the 1st layer combination is based on redundant information, different modes for the combination exist depending on the available data. Nevertheless, the combination is always realized with Eq. (4) to Eq. (6) independent of the information the layers can provide. To this end, four different modes can be distinguished: all layers can provide information, thereby the lane information from the L_{eH} layer can be refined by the L_{MR} layer and is underlined by the L_{UR} layer. Second, only the L_{eH} layer together with the L_{UR} layer have data, providing information about lanes and directions but not with the precision of the L_{MR} layer. Third, only the L_{MR} layer together with the L_{UR} layer provide information, where the driving directions of the lanes can only be inferred from other sources (e.g. the current scene context). Finally, just the L_{UR} layer provide information about the unmarked road, which is the basis information.

The combination of the L_{UR} and L_{eH} layer is a simple multiplication where L_{eH} is reduced to a binary matrix of the desired lane(s). If the information is not available the resulting matrix is just filled with ones having no impact on further processing steps.

The marked road layer is so far shown as one layer, actually it is divided in six sub-layers corresponding to three lane markings to the left (M_i^L) and three to the right (M_i^R) of the current ego position (with $i = \{1, 2, 3\}$). Additionally, not only the position of the road marker for each sub-layer M_i^D is set to one in the sub-layer, but also the area left of a lane marker to the left and right of a lane marker to the right. This will generate a mask for further processing. In mathematical terms a function providing only the lane markers ($om_i^D(x, y)$) is checked for all points p that satisfy the corresponding condition, which is given by:

$$\forall p \leq x \text{ with } om_i^L(x, y) = 1 \text{ is } m_i^L(p, y) = 1 \quad (2)$$

$$\forall p \geq x \text{ with } om_i^R(x, y) = 1 \text{ is } m_i^R(p, y) = 1. \quad (3)$$

Hence, the following lanes can be extracted, the ego lane (Lane_{own}):

$$L_1^{\text{com}}(\text{Lane}_{\text{own}}) = (L_{UR} \cdot L_{eH}(\text{Lane}_{\text{own}})) - M_1^L - M_1^R \quad (4)$$

the first (Lane_1^D) and second (Lane_2^D) lane to the left and right:

$$L_1^{\text{com}}(\text{Lane}_1^D) = (L_{UR} \cdot L_{eH}(\text{Lane}_1^D) \cdot M_1^D) - M_2^D \quad (5)$$

$$L_1^{\text{com}}(\text{Lane}_2^D) = (L_{UR} \cdot L_{eH}(\text{Lane}_2^D) \cdot M_2^D) - M_3^D \quad (6)$$

$$\text{with } D \in \{\text{L}, \text{R}\}.$$

However, it is also possible to extract a number of adjoining lanes at the same time, by changing M_i^D to the outmost left and right lane marker of the adjoining lanes in Eq. (5). If available the eHorizon provides additionally the driving direction of the extracted lane(s).

The attention system can also be modulated by the so provided lane information. For example, the search for car-like openings on the road can be restricted to certain lanes, a number of lanes and also the overall road. This allows a specific focus on relevant areas of the surrounding environment for the attention, e.g. only the oncoming traffic lane can be focused, since there is the highest probability for emerging new traffic participants. If the eHorizon information is not available, the driving direction of the lanes can also be inferred from the scene context, assuming the same driving direction for all lanes on highways and opposing driving direction on the left lanes in inner-city and rural roads.

5. Example Task

Depending on the current task the combination of the i layers L_i^{com} is performed. In order to make this point clear, an example task is carried out illustrating the concept.

As task the computation of a *possible stop position* is given (also illustrated in Fig. 5). The first layer combination L_1^{com} is already described above and has the sub-task of extracting the ego-lane (Eq. (4)).

In the following stage, L_1^{com} has to be combined with the traffic sign layer L_{TS}. Hence, only the relevant traffic signs (for this task *Stop* and *Give Way*) will be kept ($L_{TS}^{\text{Stop,GW}}$). So far, each traffic sign occupies a single cell of the layer, but in order to provide a virtual limit line, the dimensions of the signs are stretched to the complete width and 1 m in depth of the $L_{TS}^{\text{Stop,GW}}$ layer. The next step is the product computation of L_1^{com} and $L_{TS}^{\text{Stop,GW}}$:

$$L_2^{\text{com}} = L_1^{\text{com}} \cdot L_{TS}^{\text{Stop,GW}}. \tag{7}$$

The result L_2^{com} is a stop position within the ego lane based on the traffic sign position. Until now, no horizontal lane markings are processed, which would deliver additional information about the stop line. But it is planned for the future to extract the "real" stop line from the environment. Please note that based on the generic attention system, a mere change in the parameters w_i^{TD} would allow the boosting of lane markers of this specific orientation. Nevertheless, in the example stream (see Fig. 6) it would anyway not be possible, due to the occlusion from the car in front.

The final step is the incorporation of the object layer. This is done similarly as Eq. (7), by substitution of $L_{TS}^{\text{Stop,GW}}$ with the object layer L_O. The result L_3^{ego} only contains (if any exist) objects on the ego lane. For all remaining objects on L_3^{ego} the distance (based on our current trajectory) is compared to the stop line (L_2^{com}) and if the object is closer, the stop line is shifted to the position of the object. The result is a spatial representation L_3^{com}, that contains the closest stop position on the ego lane.

Therefore, the task *find possible stop position* is fulfilled. As stated before, due to our generic system design, many other tasks are supported, e.g. extract objects on certain lanes (overtaking, lane change, turning lane, etc.), find corresponding maximum speed of a lane (highway with different speeds for lanes), extract ego lane for left/right turn, handle complex crossroads and so on. The important aspect is that for many new tasks the information is already available and only the layers have to be combined in a different way. Some tasks require new STM layers with new information, but even these can be easily incorporated. To this end, the generic nature is not the variation of the representations itself, but the simple change of the content within the representations with each task.

IV. Results

In Section IV-A we will provide references to evaluations of different individual system modules that play the most important role in our cognitive ADAS architecture. In Section IV-B the overall system properties of the task-dependent representation generation will be assessed. Results for different tasks are shown based on two inner-city scenarios. Additionally, the performance gain on the system level is exemplarily shown on the traffic sign classification.

A. Evaluation of System Modules

The results presented in [1] support the generic nature of the TD-tuneable attention subsystem during object search. Following this concept, the task-specific tunable attention system can be used for scene decomposition and analysis, as it is shown exemplarily on the inner-city scene in Fig. 3.

Moreover, we see the attention system as a common tuneable front-end for various other system tasks, e.g., for lane marking detection (see Section III-B). For an evaluation of the lane marking detection module, please refer to [2].

An extensive performance evaluation of the unmarked road detection can be found in [26]. Please refer to [20] for a detailed evaluation of the traffic sign recognition with an array of weak classifiers. Results of the visual scene classification can be found in [30].

B. Evaluation of Overall System Performance

In order to show the performance on the system level the classification of traffic signs (*Stop* signs) is chosen as example. A number of different image sequences with and without *Stop* signs are the basis for the evaluation. More specifically, the evaluation is based on 1569 frames in total with 399 frames showing relevant *Stop* signs,

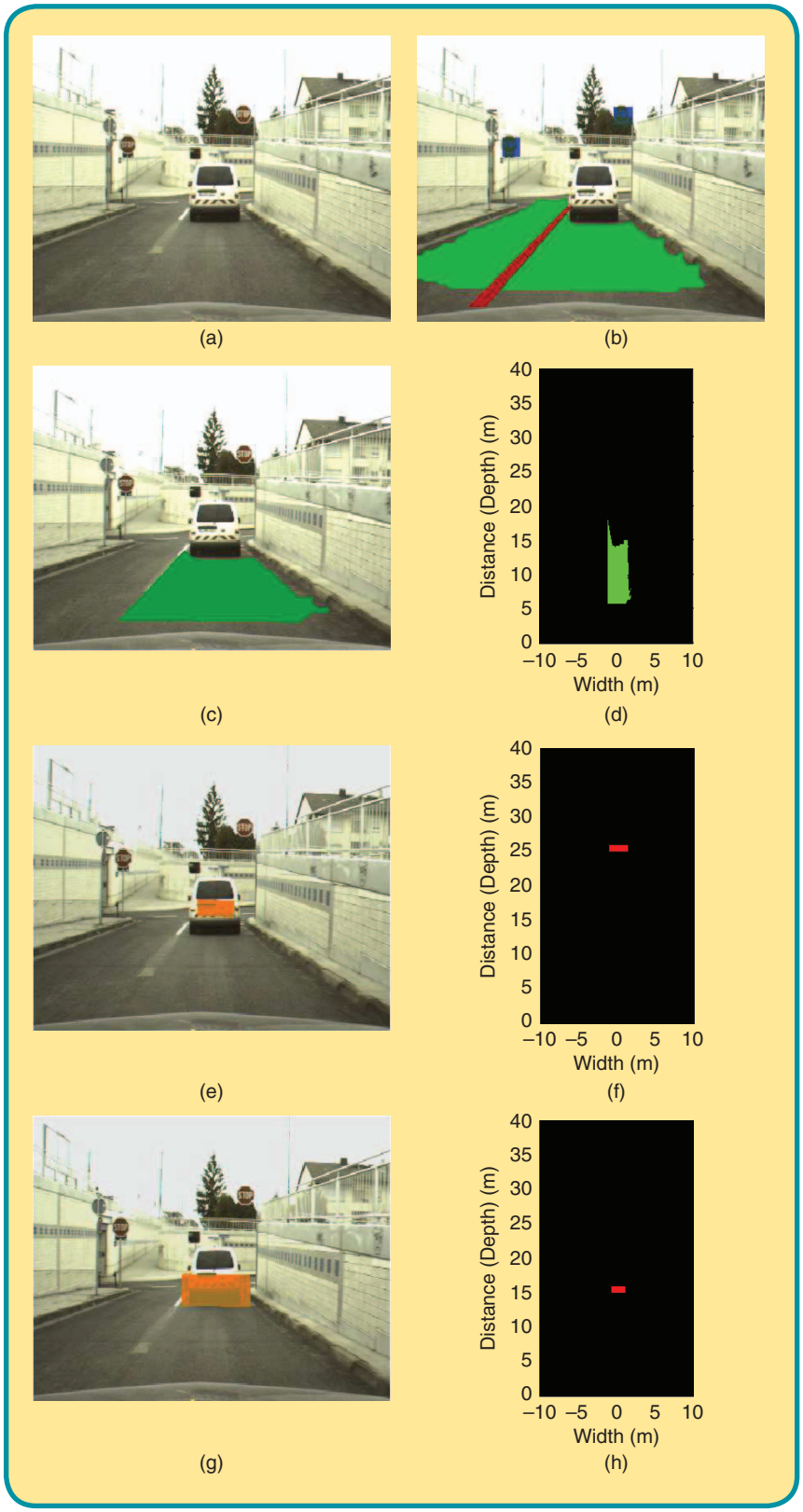


FIG 6 Second qualitative example; visualization of different results for the test stream: (a) Input image, (b) Results of the different processing modules, (c) Result image for L_1^{com}, (d) Spatial representation for L_1^{com}, (e) Result image for L_2^{com}, (f) Spatial representation for L_2^{com}, (g) Result image for L_3^{com}, (h) Spatial representation for L_3^{com}.

which are 14 different sequences with 8 showing relevant *Stop* signs. In general, two different evaluations will be done in the following. First, the temporal coherence of the image sequences is neglected. Therefore, it is required to find every traffic sign on each image, no matter if it is the same physical sign from an image before and also the same in the next image. The described evaluation type assesses the so-called single image performance (which is usually done to show the performance of a classification). The second evaluation type is done on physical traffic signs. Therefore, each physical traffic sign has to be found only once, no matter for how many frames it is visible. This evaluation is better suited for an ADAS, since the focus can be shifted to find all physical signs with the lowest false positive (FP) rate. The false positive detections can cause false alarms of the assistance system and therefore the development process should aim at minimizing this error type. The Receiver Operating Characteristic (ROC, see [32]) plot is used to evaluate the performance of a classifier. Hence, Figure 7 shows the first evaluation on the single image performance, while Figure 8 shows the second evaluation on the physical sign performance.

In both plots the following results of system modules (and their combination) are depicted: only the sign classification by the Array of Weak Classifiers (AWC), the combination of the AWC with the eHorizon (eH), the generic temporal integration (Ti) of the AWC and the combination of all modules (AWC + Ti + eH). The resulting ROC curves show a separation between results with and without temporal integration. However, the integration of the eHorizon always increases the performance and the temporal integration shows a significantly lower FP rate for similar Recall rates.

As the ROC in Figure 7 shows, the single image performance is decreasing for all combinations that include

the temporal integration. Due to the required (repeated) detection in consecutive images (the temporal integration is done over three frames) the first two classifications can not lead to a positive recognition. In order to change this behaviour other temporal integration procedures can be used (e.g. voting). Nevertheless, the single image performance is rather unimportant on the system level. It is much more important that all physical traffic signs are at least detected once. Therefore, Figure 8 shows the ROC plot based on physical traffic signs. In order to show the differences the axis of abscissa has a logarithmic scale. All physical traffic signs are found by each of the combinations, but with different false positive rates. The best results are shown for the full system combination (AWC + Ti + eH). Based on that a more profound analysis becomes necessary in order to explain the strong variations between the single image performance in contrast to the performance on physical traffic signs. It turns out that the detection distance is a key factor for explaining the gathered evaluation data.

Table 1 shows the gathered results for the different combinations of system modules. The first visibility of a traffic sign and the corresponding distance show a variation between 23 m and 48 m for the used image sequences (the same sequences as before). Hence, the mean detection distance provides an indication on how early a system module combination can find the traffic signs. At that point, the difference between the combinations with and without temporal integration is resolved, the temporal integration reduces the mean detection distance by a few meters. Additionally, it is important to compare the FP/frame for the different combinations and at that point it becomes obvious that the system approach with AWC, Ti and eH has the best ratio between FP/frame and mean detection distance. In addition it has to be mentioned that the detection distance is correlated with the image size (in our case 300x400 pixel). If the resolution is increased also the detection distance will increase.

In order to qualitatively evaluate the presented task-dependent representation generation the intermediate representations of two inner-city scenes are visualized. The first one shows the processing results for different tasks only based on the unmarked road layer and eHorizon layer (see Figure 4). In the example no road markings are available and the traffic signs are too far away to be visually recognizable. Nevertheless, the task is to extract the ego-lane, the opposing lane and restrictions on the ego-lane imposed by the existing traffic signs. To this end, one task fulfilled by the first layer combination L_1^{com} is the extraction of the ego-lane shown in green on Figure 4a. Another task of L_1^{com} is the extraction of the opposing lane shown in red (see Figure 4a). Extracting the opposing lane is possible even in the absence of lane markers, based on the eHorizon data. Due to the *No Entry* traffic sign located

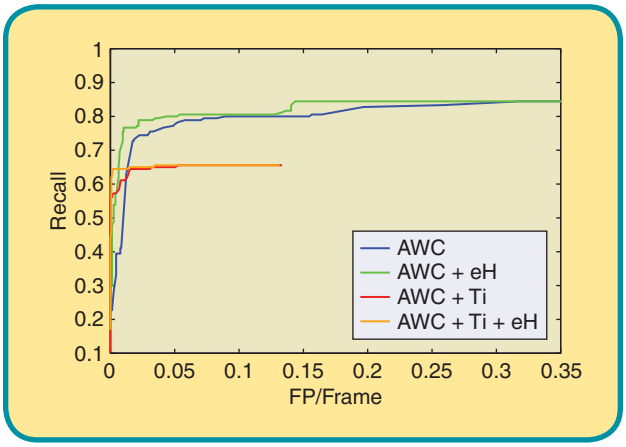


FIG 7 ROC plot showing the single image performance for the different combinations of system modules.

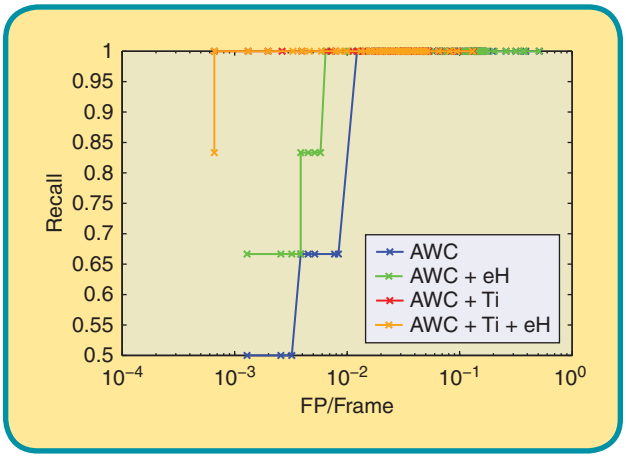


FIG 8 ROC plot showing the performance on physical traffic signs for the different combinations of system modules.

beyond the intersection, a driving restriction on the ego-lane results in the test scenario. The eHorizon provides the mentioned information as a virtual sensor in this case, hence the computation of the 2nd layer combination gets possible even with an restricting traffic sign beyond the camera range. The result of L_2^{com} is shown as visual barrier of 1 m height in Figure 4a highlighting the *No Entry* situation related to the current driving direction. In Figure 7b, the result of the unmarked road detection is visualized.

Table 1. Evaluation of the traffic sign recognition performance for the different combinations of system modules.

Combination	Mean Detection Distance	FP/Frame
AWC	35.7 m	0.0618
AWC + eH	36.0 m	0.0261
AWC + Ti	28.3 m	0.0090
AWC + Ti + eH	30.7 m	**0.0013**

The digital map/eHorizon for the current timestep is visualized in Figure 4c.

In the next example (see Fig. 6), also the internal layers are shown to give a better insight. Furthermore, results in form of 4 consecutive sample frames of a test stream are presented that show a complex real-world scenario. For the four consecutive images the results for different layer combinations are depicted. At first, an image without any annotations is shown in Fig. 6a, the results of the system processing modules are shown in Fig. 6b (red: marked road, green: unmarked road, blue: traffic signs). In the following, the results for the different layers are shown. Starting with the 1st layer combination (L_1^{com}) with the task of ego-lane extraction. Therefore, the spatial representation (Fig. 6d) shows the ego-lane in metric coordinates as extracted by the combination of the unmarked and marked road detection. Additionally, Fig. 6c shows the back-projection of the ego-lane to the image. Followed by the 2nd layer combination (L_2^{com}), having the task of limit line extraction for the detected stop line. To this end, the spatial representation (Fig. 6f) depicts the stop line taking the ego-lane and the position of the *Stop* sign into account. The stop line was back-projected to the image with a height of 1 m acting as a virtual wall (see Fig. 6e). Finally, the 3rd layer combination (L_3^{com}) is shown (see Fig. 6g-h) and therewith the task of extracting the limit line under consideration of other objects. Therefore, the detected car on our ego-lane shifts the limit line closer, again depicted as a virtual wall of 1 m height.

V. Summary and Outlook

In this contribution, we presented a novel way of scene analysis based on spatial representations. The approach is able to deal with heterogeneous processing results as input, is easily extendable with new input results, and allows a straightforward realization of tasks by a mere combination of the layers. Additionally, the scene analysis is done task-specifically, only extracting the spatial information which is currently required. The incorporation of digital map data allows a further spatial prediction horizon, as well as new fusion possibilities for the task-dependent representation generation.

The task-based environment representation is embedded in an integrated, advanced driver assistance system that relies on human-like cognitive processing principles. The system uses a biologically motivated attention system as flexible and generic front-end for all visual processing. Based on top-down links modulating the attention task-dependently, a state-of-the-art object classifier, the array of weak classifiers for traffic sign classification, a road recognition and a scene classification, we realized a highly flexible and robust system architecture.

In the future, we plan a prediction for the next n timesteps, based on a certain task. For example, the task of extracting objects to lanes would not only show the current spatial relation, but also the predicted movement. Therefore, a car on the right lane with a left indicator light will be predicted on the ego-lane. The prediction allows a preparation of the brake, if the car changes to our lane, proactively keeping a safe distance to the preceding traffic.

About the Authors

Robert Kastner received the Dipl.-Ing. degree in electrical engineering and information technology from Darmstadt University of Technology, Darmstadt, Germany, in 2007. Until December 2010 he was working on his Ph.D. degree in the Control Theory and Robotics Laboratory in cooperation with the Honda Research Institute Europe GmbH, Offenbach, Germany. The author is now working in the Functions Technology Group for the Honda R&D Europe (Deutschland) GmbH. His interests include visual scene analysis, estimation of object motion, spatial environment representations, and system integration.

Thomas Michalke received the Diploma degree in industrial engineering in 2006 and the Ph.D. degree in electrical engineering and information processing in 2009 both from Darmstadt University of Technology, Germany. His Ph.D. project was carried out at the Control Theory and Robotics Lab with the Honda Research Institute Europe GmbH, Offenbach, Germany as industrial partner. The author is currently working as a research engineer for the Daimler AG with focus on collision avoidance and mitigation. His research interests cover vision-based road and object recognition, sensor fusion, stereo vision, large scale system design, and real-time optical flow computation.

Jürgen Adamy received the Dipl.-Ing. degree in electrical engineering and the Ph.D. degree in control theory from the University of Dortmund, Dortmund, Germany, in 1987 and 1991, respectively. From 1991 to 1995, he worked as a Research Engineer, and from 1995 to 1998, as a Research Manager, at the Siemens Research Center, Erlangen, Germany, where he was responsible for the development of controls for the paper industry, steel industry, and hydraulic systems, as well as mobile robots. Since 1998, he has been a Professor at the Darmstadt

University of Technology, Darmstadt, Germany, and Head of the Control Theory and Robotics Laboratory. He is also currently Executive Director of the Institute of Automatic Control at Darmstadt University of Technology.

Jannik Fritsch received the Dipl.-Ing. degree in electrical engineering from Ruhr-University Bochum in 1996. In 1998 he joined the Applied Computer Science group at Bielefeld University where he received the Ph.D. degree in 2003. In 2004 he joined the EU Project COGNIRON (The Cognitive Robot Companion) where he headed the integration efforts for the Key Experiment Robot Home-Tour. Since 2006 he is working as Principal Scientist at the Honda Research Institute Europe GmbH in Offenbach, Germany, in the "Attentive Co-Pilot" project. His research interests are image processing methods for environment perception, spatial representations, and cognitive system concepts for intelligent automotive systems.

Christian Goerick received the Diploma degree in electrical engineering and the Ph.D. degree in electrical engineering and information processing from Ruhr-Universität Bochum, Bochum, Germany. During his time in Bochum, he was Research Assistant, DoctoralWorker, Project Leader, and Lecturer in the Institute for Neural Computation and Chair for Theoretical Biology. The research was concerned with biologically motivated computer vision for autonomous systems and learning theory of neural networks. He is currently a Chief Scientist with the Honda Research Institute Europe GmbH, Offenbach, Germany. His research interests are behavior-based vision, audition, behavior generation, cognitive robotics, advanced driver assistance systems, system architecture, and hard- and software environments.

References

[1] 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," in *Proc. IEEE Intelligent Vehicles Symp.*, 2008.

[2] T. Michalke, R. Kastner, J. Fritsch, and C. Goerick, "Towards a proactive biologically-inspired advanced driver assistance system," in *Proc. IEEE Intelligent Vehicles Symp.*, 2009.

[3] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, vol. 22, no. 6, pp. 46–57, June 1989.

[4] T. Nguyen, M. Meinecke, M. Tornow, and B. Michaelis, "Optimized grid-based environment perception in advanced driver assistance systems," in *Proc. IEEE Intelligent Vehicles Symp.*, 2009.

[5] C. L. Colby, "Action-oriented spatial reference frames in cortex," *Neuron*, vol. 20, no. 1, pp. 15–24, 1998.

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

[7] European Commission Information Society. (2007). Intelligent car initiative. [Online]. Available: http://ec.europa.eu/informationsociety/activities/intelligentcar

[8] A. Amditis, E. Bertolazzi, M. Bimpas, F. Biral, P. Bosetti, M. Da Lio, L. Danielsson, A. Gallione, H. Lind, A. Saroldi, and A. Sjoegren, "A holistic approach to the integration of safety applications: The insafes subproject within the European framework programme 6 integrating project prevent," *IEEE Trans. Intell. Transport. Syst.*, vol. 11, no. 3, pp. 554–566, 2010.

[9] H. Loose, U. Franke, and C. Stiller, "Kalman particle filter for lane recognition on rural roads," in *Proc. IEEE Intelligent Vehicles Symp.*, 2009.

[10] N. Fairfield and D. Wettergreen, "Evidence grid-based methods for 3d map matching," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2009.

[11] P. Trahanias, W. Burgard, A. Argyros, D. Hahnel, H. Baltzakis, P. Pfaff, and C. Stachniss, "Tourbot and webfair: Web-operated mobile robots for tele-presence in populated exhibitions," *IEEE Robot. Automat. Mag.*, vol. 12, no. 2, pp. 77–89, 2005.

[12] J. Knaup and K. Homeier, "Roadgraph—Graph based environmental modelling and function independent situation analysis for driver assistance systems," in *Proc. IEEE Int. Conf. Intelligent Transportation Systems (ITSC)*, 2010.

[13] R. Gregor, M. Lutzeler, M. Pellkofer, K.-H. Siedersberger, and E. Dickmanns, "EMS-vision: A perceptual system for autonomous vehicles," *IEEE Trans. Intell. Transport. Syst.*, vol. 3, no. 1, pp. 48–59, Mar. 2002.

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

[15] T. Michalke, R. Kastner, J. Adamy, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko, A. Gepperth, J. Fritsch, and C. Goerick, "An attention-based system approach for scene analysis in driver assistance," *Automatisierungstechnik*, vol. 56, no. 11, pp. 575–584, 2008.

[16] S. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[17] S. Frintrop, "Vocus: A visual attention system for object detection and goal-directed search," Ph.D. dissertation, Univ. Bonn, Germany, 2006.

[18] T. Michalke, "Task-dependent scene interpretation in driver assistance," Ph.D. dissertation, Tech. Univ. Darmstadt, Germany, 2010.

[19] H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant object recognition," *Neural Comput.*, vol. 15, no. 2, pp. 1559–1588, 2003.

[20] R. Kastner, T. Michalke, T. Burbach, J. Fritsch, and C. Goerick, "Attention-based traffic sign recognition with an array of weak classifiers," in *Proc. IEEE Intell. Vehicles Symp.*, 2010.

[21] P. Viola and M. Jones, "Robust real-time object detection," *Int. J. Comput. Vis.*, 2001.

[22] M. Bertozzi and A. Broggi, "Vision-based vehicle guidance," *Computer*, vol. 30, no. 7, pp. 49–55, July 1997.

[23] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Trans. Intell. Transport. Syst.*, vol. 10, no. 4, pp. 572–583, 2009.

[24] O. Ramstroem and H. Christensen, "A method for following unmarked roads," in *Proc. IEEE Intelligent Vehicles Symp.*, 2005, pp. 650–655.

[25] E. Dickmanns and B. Mysliwetz, "Recursive 3-d road and relative ego-state recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 2, pp. 199–213, 1992.

[26] T. Michalke, R. Kastner, M. Herbert, J. Fritsch, and C. Goerick, "Adaptive multi-cue fusion for robust detection of unmarked inner-city streets," in *Proc. IEEE Intelligent Vehicles Symp.*, 2009.

[27] V. Blervaque, K. Mezger, L. Beuk, and J. Loewenau, "ADAS Horizon: How digital maps can contribute to road safety," in *Advanced Microsystems for Automotive Applications 2006 (VDI-Buch)*, J. Valldorf and W. Gessner, Eds. Berlin: Springer-Verlag, 2006, pp. 427–436.

[28] O. Pink and C. Stiller, "Automated map generation from aerial images for precise vehicle localization," in *Proc. IEEE Int. Conf. Intelligent Transportation Systems (ITSC)*, 2010.

[29] M. E. Najjar and P. Bonnifait, "Road selection using multicriteria fusion for the road-matching problem," *IEEE Trans. Intell. Transport. Syst.*, vol. 8, pp. 279–291, 2007.

[30] R. Kastner, F. Schneider, T. Michalke, J. Fritsch, and C. Goerick, "Image-based classification of driving scenes by hierarchical principal component classification (HPCC)," in *Proc. IEEE Intelligent Vehicles Symp.*, 2009.

[31] R. M. Klein, "Inhibition of return," *Trends Cogn. Sci.*, vol. 4, no. 4, pp. 138–145, Apr. 2000.

[32] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, pp. 1145–1159, 1997.

ITS