Towards a Task Dependent Representation Generation for Scene Analysis

Robert Kastner, Thomas Michalke, Jannik Fritsch, Christian Goerick

2010

Preprint:

This is an accepted article published in IEEE Intelligent Vehicles Symposium (IV). The final authenticated version is available online at: https://doi.org/[DOI not available]

Towards a task dependent representation generation for scene analysis

Robert Kastner*, Thomas Michalke[1], Jannik Fritsch◇, Christian Goerick◇

*Darmstadt University of Technology
Institute for Automatic Control
D-64283 Darmstadt, Germany
{robert.kastner}
@rtr.tu-darmstadt.de

◇Honda Research Institute Europe GmbH
D-63073 Offenbach, Germany
{jannik.fritsch,
christian.goerick}
@honda-ri.de

***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 the novel task dependent generation of spatial representations, allowing task specific extraction of knowledge from the environment based on our biologically motivated ADAS. Additionally, the hierarchy of the approach provides advantages when dealing with heterogeneous processing modules, a large number of tasks and additional new input cues. First results show the reliability of the approach.**

Keywords: driver assistance, scene analysis, environment representation

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 specialised 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 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.

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. Therefore, biological vision systems 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 as 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, meaning the capability of a system to actively decide based on the current system state and sensor input, which task to attend 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. To this end, in the scientific community the field of designing and researching spatial representations has gained interest in the recent years. In most of the related research some kind of evidence 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 e.g. moving objects, cars, etc. from an occupancy grid (see [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 e.g. 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, exploiting spatial relations on a higher level of abstraction.

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 approach that is able to integrate different types of processing results in a generic way. With such an approach

[1]After finishing his PhD in cooperation between Honda Research Institute Europe GmbH and Darmstadt University of Technology, the author now works for Daimler AG (email: thomas_paul.michalke@daimler.com).

the extraction of information on a higher level of abstraction becomes possible. The proposed system is able to deal with complex scenes and generate spatial expectations for the current task. The realized system is tested on real-world data and first results are shown.

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.

Publications that deal in general with spatial representations are quite numerous. Nevertheless, a lot of these contributions use an evidence grid to integrate sensor data over time (see e.g. [8]). An evidence grid also provides a framework for a probability based approach, since 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 e.g. [9]. 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 [10], proposing a method for distinguishing between static and dynamic objects when building an environment map. To this end, the focus of publications regarding evidence grids is mainly on sensor fusion, temporal integration and knowledge extraction from sensor data, in contrast to our work which allows a combination of results from different heterogeneous processing modules at later processing levels. Therefore, 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.

In terms of complete vision systems, there have been a number of publications concerning the topic, please refer to [11] for a detailed comparison. One of the most prominent examples is a system developed in the group of E. Dickmanns [12]. 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.

A vision system approach in the vehicle domain that also includes an attention system and that hence is somewhat related to the here presented ADAS is described in [13]. 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.

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 with task dependent scene analysis is depicted in Fig. 1. It consists of four major parts: the "what" pathway, the "where" pathway, a part executing "static domain specific tasks" and a part allowing "environmental interaction".

The distinction between "what" and "where" processing path is somewhat similar to the human visual system where the dorsal and ventral pathway are typically associated with these two functions (see, e.g. [14]). 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 [14]). 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 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 [11]). In Section III-D, the task dependent representation generation is explained in detail.

A. The "what" pathway

Starting in the "what" pathway the 400x300 pixel color input image is analyzed by calculating the saliency map S^{total}. The saliency map S^{total} 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 [15] 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 (see Fig. 2 for a visualization). 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 [1].

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 [16]. The class of traffic signs is treated separately with an

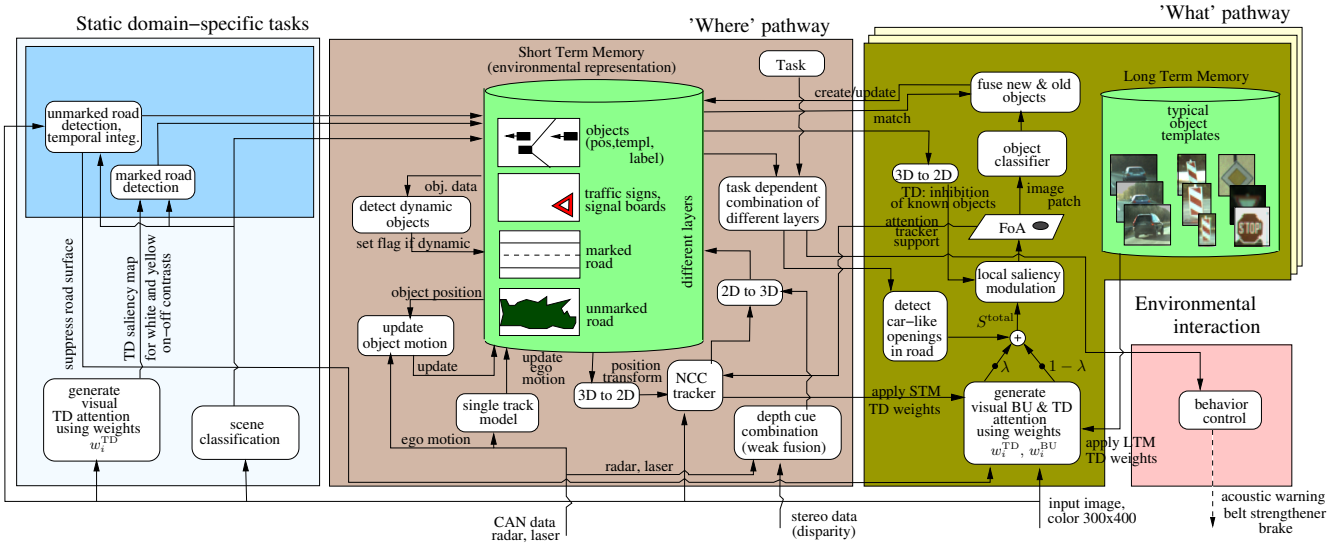

Fig. 1. Biologically motivated system structure for task dependent scene analysis using spatial representations.

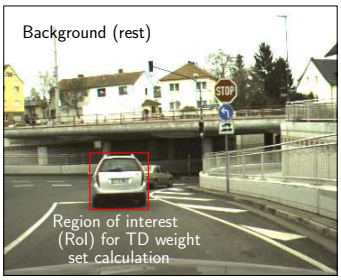

Fig. 2. Object region (RoI) for w_i^{TD} calculation against the background.

array of weak classifiers for classification as described in [17]. This procedure (attention generation, FoA segmentation and classification) models the saccadic eye movements of mammals, where a complex scene is scanned and decomposed by sequential focusing of objects in the central 2-3° foveal retina area of the visual field.

Internal information fusion processes improve the performance of system modules. For example, 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. This car-like openings are detected by searching for car-sized openings in the road segment (see [11] 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).

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 $m_{RoI,i}$ for all supported 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 [1]). 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. 1 as multiple "what" pathways). In the default case, a specific "what" pathway searches for a generic LTM object class. This is done by computing the geometric mean of all TD weight sets of the LTM object class.

B. Static domain specific tasks

In the following part, the domain specific tasks are described. These are on the one hand related to marked and unmarked lane detection and on the other hand a reliable scene classification. 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 [18] for details) before applying the Hough transform. Therefore, a clothoid model-based approach for detecting the markings is used (see, e.g., [19], [20], [21] for related clothoid based approaches). But with the knowledge of the current scene context (see later in this section) a prior for the scene specific lane width is set for the evaluation of the Hough space (e.g. for highways a lane width of around 3.7m is expected). To this end, the result of the marked lane detection is directly in metric coordinates suitable for a task dependent representation.

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 [22] for a comprehensive description of the overall procedure). The result of the unmarked road detection is also in metric coordinates.

The final part of the static domain specific tasks is the state-of-the-art 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 parameterise 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 an 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 off all parts. For more information please refer to [23].

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 (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 and fusion 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.

Starting with the former, the STM contains different layers which are used to store different classes (see Fig. 1, STM within the "where" pathway), therefore the update/fusion process is strongly simplified, if only having to cope with elements of the same class. 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 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.

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 the newly detected object 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 then compared to the new object O_{new} by means of the distance measure $\delta(O_f, O_{new})$ 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 $H_i^{O_{new}}$ (see Eq. (1)).

$$\delta(O_f, O_{new}) = \sum_{i=1}^{N} \sqrt{1 - \gamma(H_i^{O_f}, H_i^{O_{new}})} \qquad (1)$$

$$\gamma(H_i^{O_f}, H_i^{O_{new}}) = \sum_{\forall x,y} \sqrt{H_i^{O_f}(x,y) H_i^{O_{new}}(x,y)}$$

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). The 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 known objects was proved to exist in the human vision system and is termed inhibition of return (IoR), refer to [24] for details.

All known objects and traffic signs are tracked using a 2D tracker that is based on normalized cross correlation (NCC). The tracker gets its anchor (i.e., the 2D pixel position where the correlation-based 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.

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 [11] 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 processing in the specific "what" pathway and searches for the lost STM object in the following frames. This is realized by calculating a

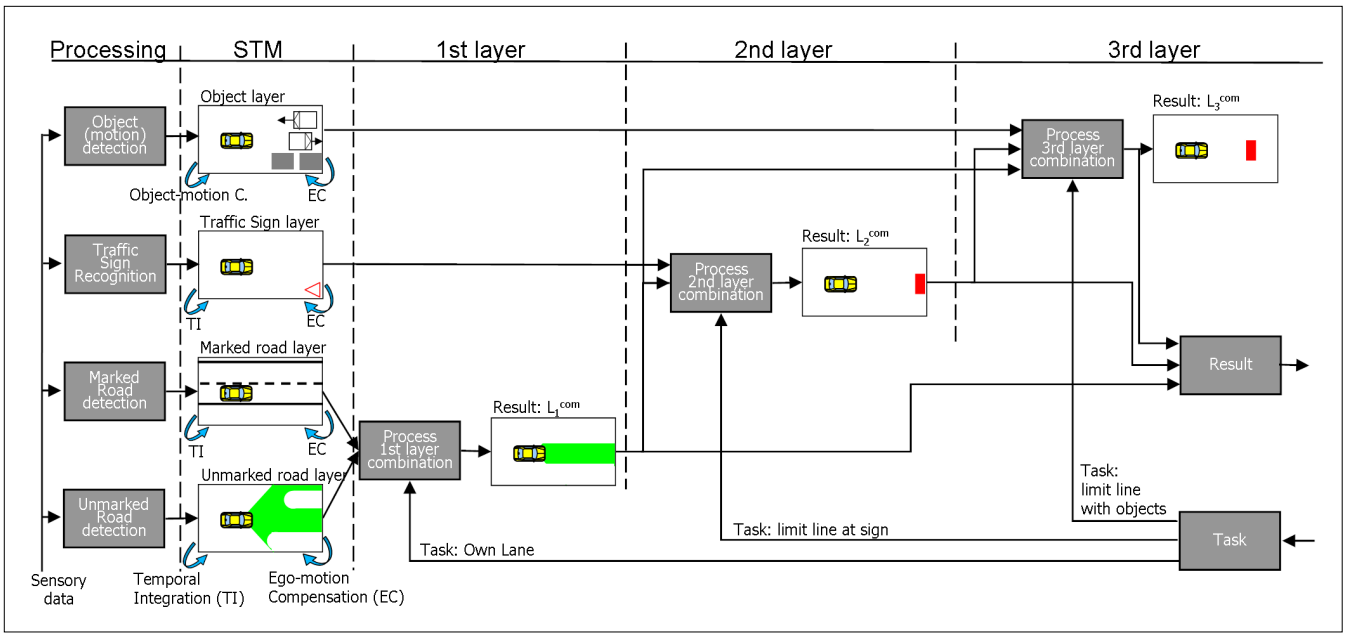

Fig. 3. Concept of task dependent representation generation.

TD weight set that is specific to the lost STM object. The object O_f found by the STM search is then compared to the searched object O_s by means of the distance measure $\delta(O_f, O_s)$ based on the Bhattacharya coefficient as already described (see Eq. (1)).

In the following, the concept of task dependent combination of different representation layers is explained. Therefore, Fig. 3 shows the strongly simplified system structure with the used hierarchy for the layer combination. The system structure of Fig. 1 is visually simplified to four processing modules providing the input for the STM on Fig. 3. Nevertheless, the functionality remains as already described. Therefore, the subsequent task dependent combination of layers is shown in more detail.

The unmarked road layer (L_{UR}) and marked road layer (L_{MR}) are always combined as 1st layer combination (L_1^{com}, see Fig. 3, 1st layer), whereas the following combinations (2nd/3rd) only depend on the current task, thus higher layers can be left out. 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 which satisfies the corresponding equation Eq. (2) or Eq. (3), generating a mask for further processing.

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

$$\forall z \geq x \text{ with } m_i^R(x, y) = 1 \text{ is } m_i^R(z, y) = 1 \tag{3}$$

Hence, the following lanes can be extracted: the ego lane (Eq. (4)), the first (Eq. (5)) and second (Eq. (6)) lane to the left and right. 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).

$$L_1^{com}(Lane_{own}) = L_{UR} - M_1^L - M_1^R \tag{4}$$

$$L_1^{com}(Lane_1^D) = (L_{UR} \cdot M_1^D) - M_2^D \tag{5}$$

$$L_1^{com}(Lane_2^D) = (L_{UR} \cdot M_2^D) - M_3^D \tag{6}$$

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

The attention system can also be modulated by the provided information, e.g. we can restrict the search for car-like openings in the road 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. So far, the driving direction of the lanes is 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.

Depending on the current task the combination of the i layers L_i^{com} is performed. To this end, an example task is carried out illustrating the concept. As task the computation of a *possible stop position* is given (also illustrated in Fig. 3). 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 ($L_{TS}^{Stop,GW}$) and their dimensions stretched from a single cell of the layer to the complete width and 1m in depth, therewith providing a limit line. The next step is the product computation of L_1^{com} and $L_{TS}^{Stop,GW}$ (see Eq. (7)), resulting in a stop position

within the ego lane based on the traffic sign position.

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

So far no horizontal lane markings are processed, which would deliver additional information about the stop line. But this is planned for the future to extract the "real" stop line from the environment. Nevertheless, in the example stream (see Fig. 5) 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 $L_{TS}^{Stop,GW}$ 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 closed stop position on the ego lane.

Therefore, the task *find possible stop position* is solved, nevertheless there are many other tasks possible, 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 thing is, that for many new tasks the information already exists 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 evaluate 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. Based on a inner-city scenario results for different tasks are shown.

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 tuneable attention system can be used for scene decomposition and analysis, as it is shown exemplarily on the inner-city scene in Fig. 4.

(a) (b) (c)

Fig. 4. Attention based scene decomposition: (a) Inner-city scene, (b) TD attention tuned to cars, (c) TD attention tuned to traffic signs

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

Because of limitations in space the extensive performance evaluation of the unmarked road detection can be found in [22]. Also, for details about the evaluation of the traffic sign recognition with an array of weak classifiers please refer to [17].

B. Evaluation of overall system performance

In order to qualitatively evaluate the presented task dependent representation generation aspects, results in form of 4 consecutive sample frames of a test stream are presented that show a complex real-world scenario (see Fig. 5). In order to show the results of the different layer combinations the four consecutive images depict the results for different layer combinations. Starting with a clear image of the scenario in Fig. 5a, the results of the systems processing modules are shown in Fig. 5b (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. 5d) shows in metric coordinates the ego lane extracted by the combination of the unmarked and marked road detection. Additionally, Fig. 5c 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. 5f) 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 1m acting as a virtual wall (see Fig. 5e). Finally, the 3rd layer combination (L_3^{com}) is shown (see Fig. 5g-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, again depicted as a virtual wall of 1m height.

V. SUMMARY AND OUTLOOK

In this contribution, we presented a novel way of scene analysis based on spatial representations, which is able to deal with heterogeneous processing results as input, is easyly extendable with new input results and tasks can be simply realised by a combination of the layers. Additionally, the scene analysis is done task specifically, only extracting the spatial information, which is currently required. In the future we also 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 brake preparation, if the car changes to our lane keeping a safe distance to the new car in front.

Also an integrated, advanced driver assistance system that relies on human-like cognitive processing principles is

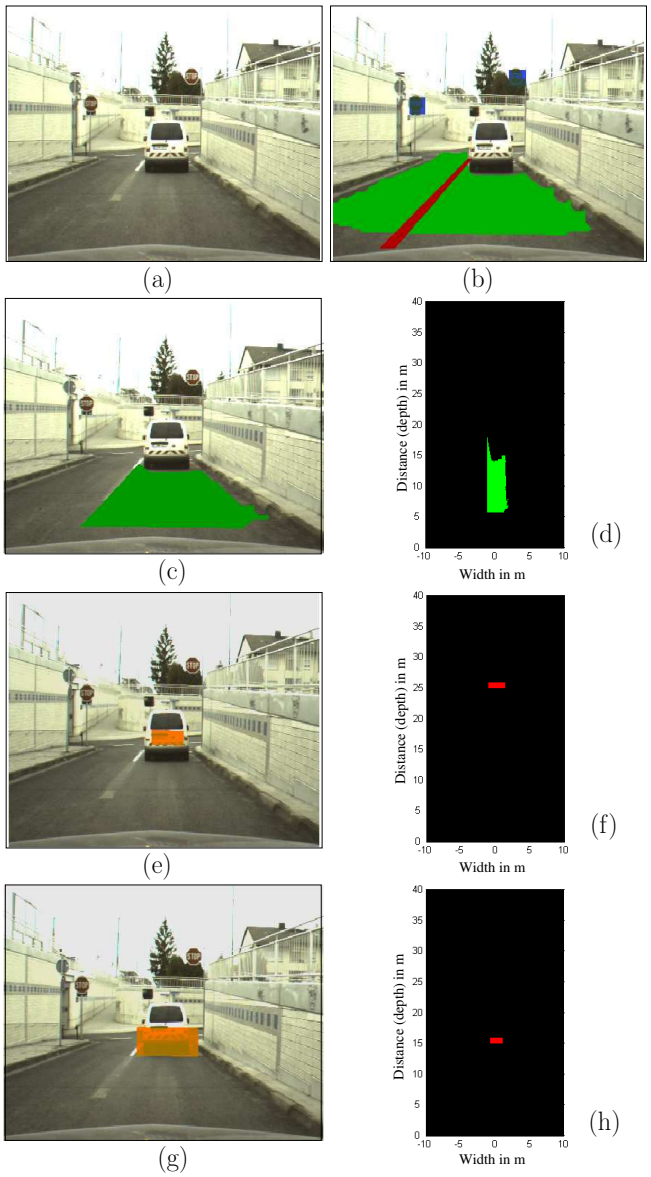

Fig. 5. Visualization of different results for the test stream: (a) Input image, (b) Results of the different processing modules, (c) Result image for L_1^{com}, (d) Spatial representation for L_1^{com}, (e) Result image for L_2^{com}, (f) Spatial representation for L_2^{com}, (g) Result image for L_3^{com}, (h) Spatial representation for L_3^{com}.

shown. 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, a road recognition and a scene classification, we realized a highly flexible and robust system architecture. We plan to port the described extensions from Matlab to C in order to integrate them in our existing online system [1] for evaluating them on our prototype vehicle.

.

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 Symposium*, 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 Symposium*, 2009.

[3] H. Moravec and A. Elfes, "High resolution maps from wide angle sonar," in *IEEE Internat. Conf. on Robotics and Automation*, 1985.

[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 Symposium*, 2009.

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

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

[7] WWW, *European commission information society 'Intelligent Car initiative*, 2007, http://ec.europa.eu/informationsociety/activities/intelligentcar/.

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

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

[10] D. Hähnel, D. Schulz, and W. Burgard, "Map building with mobile robots in populated environments," in *IEEE/RSJ Internat. Conf. on Intelligent Robots and Systems (IROS)*, 2002.

[11] T. Michalke, R. Kastner, J. Adamy, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko, A. Gepperth, J. Fritsch, and C. Goerick, *at - Automatisierungstechnik*, vol. 56, no. 11, 2008.

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

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

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

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

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

[17] 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 Intelligent Vehicles Symposium*, 2010.

[18] A. Broggi, "Robust real-time lane and road detection in critical shadow conditions," in *Proc. IEEE Internat. Symp. on Computer Vision*, Parma, 1995.

[19] U. Franke, H. Loose, and C. Knoeppel, "Lane recognition on country roads," in *Proc. IEEE Intelligent Vehicles Symposium*, 2007.

[20] O. Ramstroem and H. Christensen, "A method for following unmarked roads," in *Proc. IEEE Intelligent Vehicles Symposium*, 2005.

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

[22] 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 Symposium*, 2009.

[23] 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 Symposium*, 2009.

[24] R. M. Klein, "Inhibition of return," *Trends in Cognitive Science*, vol. 4, no. 4, pp. 138–145, April 2000.