Proceedings of the
2007 IEEE Intelligent Vehicles Symposium
Istanbul, Turkey, June 13-15, 2007

About robust hypotheses generation and object validation in traffic scenes

Martin Schneider
Technische Universität Darmstadt
Institute for Automatic Control
D-64283 Darmstadt, Germany
Email: mschneider@rtr.tu-darmstadt.de

Jens Gayko
Honda R&D Europe (Deutschland) GmbH
D-63073 Offenbach/Main, Germany
Email: jens_gayko@de.hrdeu.com

Christian Goerick
Honda Research Institute Europe GmbH
D-63073 Offenbach/Main, Germany
Email: christian.goerick@honda-ri.de

***Abstract*— Environmental perception is an important element of Advanced Driver Assistance Systems. The perception mainly consists of the steps sensing and data interpretation. Both of these steps are affected by errors due to noise and misinterpretations. Therefore, we present a system design addressing the problem of robust processing under limited resources in a hierarchical system architecture that can use state of the art data-fusion and object-recognition methods.**

I. Introduction

Advanced driver assistance systems (ADAS) require a proper interpretation of the vehicle's environment. While comfort systems are allowed to fail in certain situations, predictive safety-systems like brake-assistants have to operate reliably under a large number of traffic scenarios and physical conditions like adverse weather. Hence one key element of ADAS is a robust environmental perception.

Street traffic is regulated by numerous traffic rules and is usually restricted to artificial environments, mainly depending on the degree of development of the respective country. Nevertheless, from an environmental perception point of view, street traffic is complex with a variety of situations and innumerable variations of environmental conditions. This complexity is still a challenge for research on environmental perception.

As a consequence, today's available ADAS focus on a clearly defined subset of situations and environmental conditions. For example the environment recognition capabilities of adaptive cruise control (ACC) systems is restricted to moving objects in the predicted vehicle path. These restrictions are caused by limitations of today's available sensors, which can sense just a few parameters of the environmental conditions, as well as limitations of today's signal processing methods.

In this paper we present a hierarchical sensor fusion concept for environmental perception and address the problem of robustness under limited resources. Resource constraints affect robustness in conventional architectures, as will be discussed in section II. By guiding the expensive parts of the processing with a computational cheap information measure, the resources can be used more efficiently. Hypothesis generation in our framework is composed of a generic object-unspecific attention-stage followed by an object-specific recognition/segmentation-stage. Sensor fusion is performed on metrical and semantical information using a multiple-hypotheses object-recognition and temporal tracking procedure. We evaluated our system approach on specific traffic scenarios and observed promising results.

The approach demonstrated in this paper is closely related to ideas of Dickmanns [3], who investigated the concept of combined foveated and peripheral vision on the level of gaze-control for autonomous driving. Therefore his architecture further includes elements like mission planning and behavior decision, which go beyond the scope of this paper. We extended some of the aspects of the architecture of Dickmanns [3] relating to the concepts of attention and object tracking by integrating the available sensors on a more general basis.

The system approach of Darms et al. [2] is more focused on the fusion task itself by introducing the concept of a *virtual sensor* that can use the information of several physical sensors on demand, depending on the current perception task. They propose two levels of description of the sensory information (L1,L2) which standardize the different sensor cues for subsequent fusion.

In the EU-sponsored project ProFusion [14], a subproject of PReVENT (Preventive and Active Safety Applications), the perception task is performed within a hierarchy consisting of the elements sensor-refinement, object-refinement and situation-refinement. Here refinement is loosely defined as the underlying process that determines the output of the different stages of the hierarchy, i.e. the sensor-refinement process delivers sensor-measurements, the object-refinement delivers objects and so on. However, our definition of refinement relates to the improvement of object-hypothesis over time due to the accumulated knowledge.

A pure symbolic representation of traffic scenarios was proposed by Gerber and Nagel et al. [5]. This approach can handle complex relations of objects in a traffic scene. It is possible to generate an unambiguous and complete description of a complex traffic scene if a reliable detection of the objects can be guaranteed. Since robustness of object recognition is still a challenge for automotive applications, these approaches are not yet established for real world applications.

The paper is structured as follows: Section II addresses the robustness-resources dilemma while section III describes the key elements of the proposed perception architecture. These key elements include the hypothesis generation mechanism composed of the attention control subsystem explained in

1-4244-1068-1/07/$25.00 ©2007 IEEE.

TABLE I
OBJECT PROPERTIES AND ESTIMATION METHODS

object property	estimation method
position	tracking / segmentation
velocity, acceleration	tracking
size, dimension	segmentation/classification
semantic category	classification

subsection III-A and the object-recognition/-segmentation methods of subsection III-B. Object-validation is based on temporal stabilized metrical and semantical information and is described in subsection III-C. Section IV demonstrates the feasibility and performance of the architecture with data captured of real-world scenarios.

II. ROBUSTNESS UNDER LIMITED RESOURCES

The term *robustness* is often used for characterizing the ability of an environmental perception system to extract the relevant information under a large variability of the external environment leading to changed sensor-signals and signal-to-noise ratios. Robustness against these signal-changes can only be achieved by extracting object-specific features that are invariant over large variations of the raw signal-characteristic, or by modeling the feature-variations.

A further aspect of robustness arises due to the mentioned signal-variability at the output of components like object-detection/classification stages. Their interpretation of the measured object-features are often used for intermediate decisions and fed into other stages of the system. This is the case, for example, when strong edge-features are used for visually estimating the lane-parameters. Therefore, the components of a perception-system need to be robust against sensor noise and incorrect intermediate decisions.

The relevant information to be extracted are usually some physical states of objects and its semantic description, and mainly depend on the task of the specific ADAS-application, see e.g. [11] and [2] for a detailed overview. Table I lists the required methods for extracting the specific information. A segmentation method binds several features together and delivers a region of a sensory-subspace corresponding to one object-hypothesis. This region can be assigned a semantic category/label by classification methods like Bayesian decision or neural networks. The term tracking stands for temporal recursive state-estimation of the physical properties like position and acceleration of objects.

The estimation of the object-properties mentioned in table I is performed by fusing the measurements of different sensors. Sensor fusion can be formulated in a probabilistic framework using the Bayesian approach (e.g. [7]) by computing the a-posteriori joint-density-function

$$p(\vec{x}_{pos}, \vec{x}_{vel}, \vec{x}_{acc}, \vec{x}_{size}, \vec{x}_{type} | \vec{y}_{pos}, \vec{y}_{size}, \vec{y}_{type}) \ , \qquad (1)$$

describing the probability-density of the estimated states $\vec{x}$ depending on the measurements $\vec{y}$. We omitted the time-dependency of the estimation variables $\vec{x}_{pos}(t)$, $\vec{x}_{vel}(t)$, $\vec{x}_{acc}(t)$ and of the measurements $\vec{y}_{pos}(t)$, $\vec{y}_{size}(t)$, $\vec{y}_{type}(t)$.

In this context, robustness can be defined as the ability of determining the most probable states $\vec{x}$ based on the measurements $\vec{y}$, which can be e.g. defined as the maximum-a-posteriori (MAP) estimate of the joint-distribution of (1). Therefore, the optimal solution to the robust estimation problem in this probabilistic framework can be obtained from the joint-distribution (1). This solution can be interpreted as computing the probability of all possible interpretations of the unknown state $\vec{x}$, that fits to the measurements $\vec{y}$. Since metrical and semantical information of the real world is estimated in (1), the definition of robustness in terms of the most probable interpretation of the measurements address both influences mentioned at the beginning of this section: The variability due to the environment as well as the false detections/interpretations of the intermediate stages of the perception system.

Why is robustness affected by the amount of computational resources?

In the general case the computation of the density (1) requires solving numerical multi-dimensional integrals and is often approximated by Monte-Carlo techniques which are widely known as particle-filtering ([4]). The usually high computational costs of the approximation mainly depends on the dimension of the state space and the complexity of the corresponding measurement likelihood-function, that have to be evaluated for each particle.

Besides the pure probability-density estimation problem, the generation of the object-hypothesis has also to be taken into account. In the case of pure distance/velocity-data delivered by a LIDAR/RADAR-sensor, the computational costs for generating object hypotheses are low because of the small resolution of the sensors. In the case of visual-measurements, the generation of the object hypotheses is the most expensive operation, since object-specific features are used for the detection, segmentation and classification of objects. This is the severe computational bottle-neck in conventional architectures. For example, in the approach of Schweiger et al. [13], the whole video-image is explored for car-features like tail-lamp-circles and symmetry. Since the extraction of all possible feature combinations is generally infeasible when dealing with more than one object and different object-categories, a hierarchical processing is required for resolving this "curse of feature-dimensionality", which is described in the following sections.

III. ARCHITECTURE

In order to overcome the general robustness-resources dilemma described in the previous section, we propose a perception hierarchy as shown in figure 1 containing the following key aspects:

1) Separation of the hypothesis generation into a global object-unspecific feature extraction stage used as an attention mechanism
2) Object-specific multiple hypotheses object recognition.
3) Multiple hypotheses object tracking and validation.

In addition, different levels of object representation (coarse/fine) are used according to the required accuracy, but

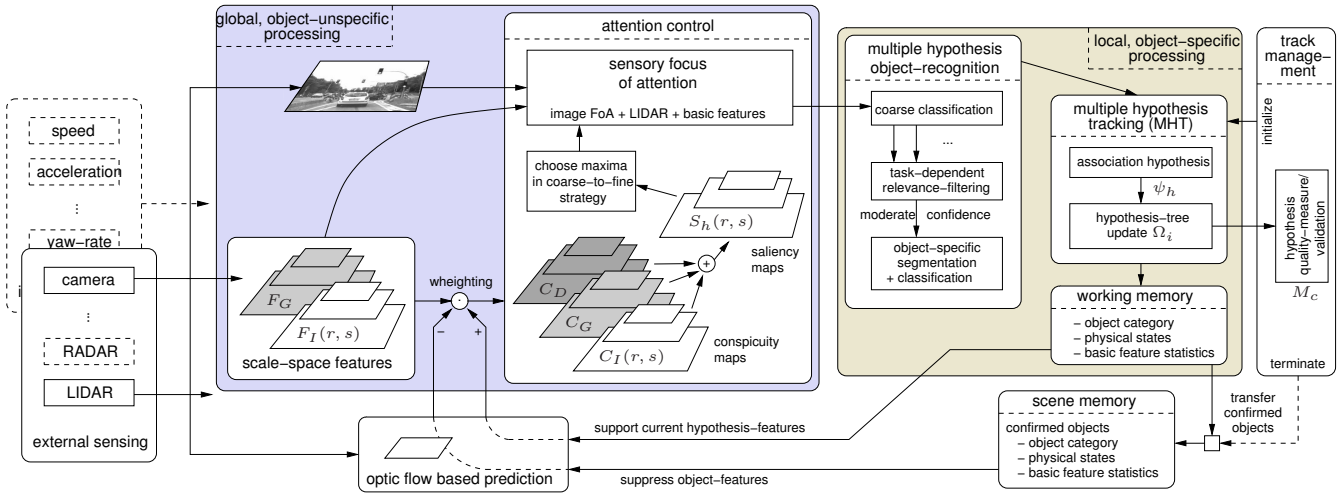

Fig. 1. System architecture

due to space-limitations this is not shown in figure 1.

Sensory information is acquired via the external sensors for measuring objects around the vehicle and the internal sensors for measuring the ego-states of the vehicle (velocity, acceleration, yaw-rate). External sensors are typical distance/velocity sensors like LIDAR and RADAR as well as image-sensors like CMOS/CCD-camera or IR-imaging sensors. These sensors cannot sense occluded objects, but RADAR for example is capable of detecting objects under bad weather conditions [6], so without clear visibility.

In order to efficiently analyze the possible maxima of the joint-distribution (1) under constraint resources, our system consists of a combined global processing and local processing loop, each acting on different spatial and temporal scales. The global processing consists of the object-unspecific feature extraction stage that guides the expensive processing stages via the attention mechanism. This mechanism selects the currently most important sensory region, called *Focus of Attention (FoA)*, that is further analyzed in the local processing loop over several time steps. Therefore, the sensory region delivered from the attention system has to be stabilized over time. This is performed via the optical flow based feature-tracking mechanism for solving the correspondence problem, see e.g. [1] for an overview of techniques. The global feature extraction and attention control system is more precisely explained in subsection III-A.

A detailed object-analysis in the local processing loop is then performed in order to formulate and validate object-trajectory-hypothesis consisting of the information mentioned in table I. An object-recognition system gives first some hints about relevant objects located in the selected region. A further object-specific segmentation step precisely locates the interesting objects in the sensory domain, which is described in subsection III-B. Based on these detailed object-analysis several object-hypothesis are generated which are subsequently evaluated by the *multiple-hypotheses-tracking (MHT)* approach that accumulates the knowledge about objects in probable trajectories over time. The MHT-method is briefly explained in subsection III-C. All components of this local processing loop are designed to produce or account for multiple hypotheses ensuring a fast convergence to the most probable object-state.

The local mechanism is closely related to the basic ideas of a fast-feedforward recognition and temporal refinement process inspired by models of the visual cortex of the human brain as described in Körner et al. [9].

During the temporal tracking-process (refinement), a coarse representation of the object is constructed based on the object-unspecific features extracted from the attention system (not shown in figure 1). After the refinement process has converged the confirmed object is transfered from the working memory to the scene-memory, where all ADAS-applications have access to it. Then the focus of attention moves to another sensory region while a coarse track-validation method is used for measuring whether the predicted states of the already confirmed objects are becoming more and more inaccurate. In that case it is necessary to update the current knowledge about the object by attending the object with the local processing loop again.

It is worth noting that this kind of coarse track-validation does not require the extraction of object-specific features, especially the computational expensive object-segmentation and classification steps are omitted. This approach results in an optimal resource management in the sense that temporal updates of the object-representations are only performed in order to keep the track.

A. Attention control system

Visual attention is a widely explored field in the psychological and biological vision-community. Wolfe et al. [16] discuss the features used for attention processes, like color, intensity, orientation, depth, motion, etc. A possible implementation of the mechanisms is proposed by Itti et al. [8], where a multi-scale filtering approach is used for extracting basic features like color-complementary, intensity and orientation. With center-surround operations the local

contrast of these features to their environment is determined and integrated into *conspicuity-maps* for each feature-modality. These conspicuity-maps are combined into a final *saliency-map* in which the visually interesting regions are quantified by a scalar value. Attention is then guided by maxima in the saliency map, while already observed regions are temporarily suppressed in order to avoid locking attention to one region in the image. This procedure ensures that the system performs *saccades* to all visually interesting locations in the image.

In our model we extended the basic idea of visual attention to the whole sensory input. We define the visual salient features as those which can be best described at a specific scale and which are maxima to their local environment. This is further explained in more detail. A scale-space representation $G(s)$ of the input-image is generated via a Gaussian pyramid. From this representation, we only use some coarse scales $s_c \dots s_{c+n}$ with $c = 4$ and $n = 2$ for further feature-extraction, since we are not interested in detailed object-recognition at this step. We rather would like to achieve a very fast generated coarse scene description. The intensity features are now computed by top-down projecting the first and second order statistics of a local image region of base-length b from a coarser scale s_{c+1} to a finer scale s_c:

$$f_I(\vec{r}_c, s_c) = 1 - e^{-\frac{(G(\vec{r}_c, s_c) - \mu_{\vec{r}_{c+1}, s_{c+1}})^2}{2\sigma^2_{\vec{r}_{c+1}, s_{c+1}}}} \tag{2}$$

The mean $\mu_{\vec{r}_{c+1}, s_{c+1}}$ and standard-deviation $\sigma_{\vec{r}_{c+1}, s_{c+1}}$ are determined in a local region of size b around the central position r_{c+1} on the upper scale s_{c+1}.

According to equation 2 the values f_I have a value near zero if they can be explained by the statistics of the corresponding feature location at the upper level. The more the value tends to 1, the better is the feature explained on the lower scale.

Similar features can be obtained from orientation-selective filters like Gabor-filters. Here we use four orientations $\varphi = 0°, 45°, 90°, 135°$ for the unbiased imaginary part of Gabor-filter

$$\begin{aligned} \mathcal{G}^{\omega}_{\varphi} &\equiv \mathcal{G}(x_1, x_2; \sigma, \omega, \varphi) = \Im\{G(x_1, x_2; \sigma, \omega, \varphi)\} \\ &= \Im\left\{ \frac{1}{2\pi\sigma^2} e^{-\frac{x_1^2 + x_2^2}{2\sigma^2}} e^{-i\omega(x_1 \cos(\varphi) + x_2 \sin(\varphi))} \right\} \\ &= -\frac{1}{2\pi\sigma^2} e^{-\frac{x_1^2 + x_2^2}{2\sigma^2}} \sin(\omega(x_1 \cos(\varphi) + x_2 \sin(\varphi))), \end{aligned} \tag{3}$$

leading to four orientation selective Gabor-feature maps $f_g(\vec{r}_c, s_c)$.

In order to also measure the strength of a feature region according to its local surrounding, we filter the resulting feature maps $f_{\cdot}(\vec{r}_c, s_c)$ with a center-surround operation. This can be a Difference of Gaussian filter (DoG) or a Laplacian of Gaussian $\nabla^2 L = L_{xx} + L_{yy}$ as proposed by Marr [10]:

$$\mathcal{L}^{\sigma} \equiv \mathcal{L}(r; \sigma) = \frac{1}{\pi\sigma^4} \left(1 - \frac{r^2}{2\sigma^2}\right) e^{-\frac{r^2}{2\sigma^2}} \ , \tag{4}$$

with $r = \sqrt{x^2 + y^2}$ as the radial distance of the center of the filter. We omitted the negative sign since we are interested in positive results. The final object-unspecific feature-maps of the intensity-features F_I and the Gabor-features F_G are obtained by

$$F_I(\vec{r}_c, s_c) = f_I(\vec{r}_c, s_c) * \mathcal{L}^{\sigma}(\vec{r}_c) \tag{5}$$

$$F_G(\vec{r}_c, s_c, \varphi) = f_G(\vec{r}_c, s_c, \varphi) * \mathcal{L}^{\sigma}(\vec{r}_c) \ . \tag{6}$$

The parameter σ is chosen to $\sigma = \frac{\sqrt{2}b}{2}$, which gives the positive kernel of the Laplacian of Gaussian-filter a radius equivalent to the base-length b.

These feature-maps are subsequently used as the bottom-up component of the feature-conspicuity-maps $C_I(\vec{r}_c, s_c)$ and $C_G(\vec{r}_c, s_c, \varphi)$, which measure the conspicuity of each feature-modality. Furthermore these maps are influenced by reinforcing those features of the current tracking-loop and inhibiting those feature-locations, that were already inspected and belong to known objects in the scene-memory (fig.1).

Further sensor cues like RADAR and LIDAR are integrated into the attention system at the level of the conspicuity-maps. For pure distance-measurements like the data from LIDAR-sensors, the sensor-beams are first clustered into groups of measurements of similar distance. A simple threshold-based clustering is used, i.e. distance measurements are bound together if their differences are below a pre-defined threshold. The clustered distance values thus represent a hypothetical object-width defined by the outermost sensor-beams of a specific cluster. From these clusters a further feature-pyramid $C_D(\vec{r}_c, s_c)$ is constructed by creating a blob of base-length b in the pyramid-level, at which the width represented by the cluster matches best the base-length. The position of the blob is determined by mapping the 3-dimensional cluster-coordinates into the Gaussian image-pyramid. The feature blobs are weighted depending on the average cluster-distance $d_{cluster}$ of each cluster as follows:

$$w_d = \begin{cases} 0 & \text{for } d_{cluster} > d_{max} \\ \frac{(d_{min} - d_{cluster})}{(d_{max} - d_{min})} + 1 & \text{for } d_{min} \le d_{cluster} \le d_{max} \\ 1 & \text{for } d_{cluster} < d_{min} \ . \end{cases} \tag{7}$$

Here, d_{min} and d_{max} are some chosen parameters for defining the slope of the weighting function. RADAR-sensors based on the Doppler-effect can further deliver the relative velocity of objects that gives a powerful cue for discriminating stationary and moving objects. In principle they can be integrated in the same manner like LIDAR-sensors.

The final saliency-maps are the weighted sum of the six feature/conspicuity maps:

$$\begin{aligned} S(\vec{r}_c, s_c) = w_I C_I(\vec{r}_c, s_c) \\ + \sum_{\varphi} w_{G_\varphi} C_G(\vec{r}_c, s_c, \varphi) + w_d C_D(\vec{r}_c, s_c) \end{aligned} \tag{8}$$

The weights are chosen with constraint to $\sum_i w_i = 1$.

For increasing those maxima that are close to the direction of the current yaw-angle of the vehicle, the saliency maps

$S(\vec{r}_c, s_c)$ are further multiplied with a heading weight-map W_h :

$$S_h(\vec{r}_c, s_c) = S(\vec{r}_c, s_c) \cdot W_h(\vec{r}_c, s_c) \ . \tag{9}$$

The heading is determined from the yaw-rate delivered by the internal sensing module.

From the saliency-representation (9) the local maxima are used as salient candidate-regions and the scale of the maxima indicate the size of the hypothetical object-region in the image domain. Scene exploration starts in a coarse-to-fine strategy, meaning that large salient regions indicated by maxima on the coarse saliency-maps are inspected first, since they might correspond to objects close to the vehicle. These regions are mapped into the original image-resolution in order to further inspect them by the high-resolution object-recognition path.

B. Object-recognition/segmentation

The delivered focus of attention (FoA) from the attention system is further analyzed for determining the object-category, its size and precise position in the image. In this section we briefly describe the local object-recognition system that use object-specific features and knowledge in order to find the most promising maximum of the joint-distribution (1).

First the object-region is fed into an object-classifier in order to get a first hint on the possible object-category. For the classification, we use the brain-like neural-network architecture proposed by Wersing et al. [15]. The network consists of four hidden layers performing feature competition, pooling and feature combination in two consecutive stages. During this processing, a feature vector $\vec{I}$ is transformed into the feature space C2, which is an established model in the biological vision community. This layer has the same order of dimension as the input image patch $\vec{I}$, thus building a high-dimensional feature space. The representation in the C2-space has the nice property, that many different object categories can be separated linearly, as shown in [15]. New object-categories can be easily trained with gradient-descend methods and only the very last stage is affected by this training. The classifier even shows a good performance according to position and size invariance of the objects, which is essentially required in our case, since we do not give a precise patch of the object-region to the classification system.

From the output of the object-classifier, we generate object-hypotheses H_{obj} for those categories that exceed a minimum recognition confidence. Based on these hypotheses an object-specific segmentation is performed that uses object-knowledge about several features. For example the segmentation of vehicle rear-views uses the symmetry as one feature as proposed by Schweiger et al. [13]. The output of the segmentation process is several bounding boxes that indicate hypothetical object-positions of a specific category in the focus of attention. The distance measurements of RADAR/LIDAR-sensors are associated to these bounding boxes if they overlap with the boxes in the image domain. This association step leads to the final object-hypotheses, $H_m(y_{type}, \vec{y}_{impos}, \vec{y}_{imsize}, \vec{y}_{dpos})$, that contain the following information:

1) object category y_{type}
2) object position in image coordinates $\vec{y}_{pos}$
3) object size in image coordinates $\vec{y}_{size}$
4) object position delivered by distance-sensors in ego-centered vehicle-coordinates $\vec{y}_d$

The first 3 items are measured by the image-processing system and the last one is delivered by a RADAR/LIDAR-sensor. The object hypotheses of the measurement space may overlap and provide the sensor-fusion-stage with various interpretations of the sensory raw-data.

C. Multiple Hypotheses Tracking

In this section the mechanism is described that finds the most probable interpretation of the object hypotheses delivered by the visual object-recognition system and the distance measurements provided by the RADAR/LIDAR-sensors.

The estimated quantities $\vec{x}_{pos} = [x_w \ y_w]^T$, $\vec{x}_{vel} = [\dot{x}_w \ \dot{y}_w]^T$, $\vec{x}_{size} = s_x = w$ describe the physical states of the object in an ego-centered coordinate frame, which are generally estimated with recursive Bayesian state-estimation techniques. The last element $s_x = w$ describes the width of the object. Here we use an Extended Kalman-Filter (EKF) for estimating the state-vector $\vec{x} = [x_w \ y_w \ \dot{x}_w \ \dot{y}_w \ w]^T$ at a sample-rate of T_s seconds with the linear system-model

$$\vec{x}_{k+1} = \Phi \vec{x}_k + \mathcal{N}(\mathbf{0}, \mathbf{Q}) \tag{10}$$

and the nonlinear measurement-model

$$\vec{y}_k = \vec{h}(\vec{x}_k) + \mathcal{N}(\mathbf{0}, \mathbf{R}); \tag{11}$$

$\vec{y} = [d_{LIDAR} \ \varphi_{LIDAR} \ x_{im} \ y_{im} \ w_{im}]^T$ is the measurement vector. $\mathcal{N}(\mathbf{0}, \mathbf{Q})$ and $\mathcal{N}(\mathbf{0}, \mathbf{R})$ are normal distributed noise-vectors, d_{LIDAR} and φ_{LIDAR} are the distance and horizontal angle delivered by the LIDAR-sensor while x_{im}, y_{im}, w_{im} are the position and width of the object in image-coordinates. The state transition matrix Φ describes a constant velocity model. The nonlinear function $\vec{h}$ of the measurement-model (11) contains the transformation of the relative position of the object into the non-cartesian LIDAR-measurement-space and the projective transformation into the pixel-based camera-coordinate-system.

A critical issue in multi-sensor data fusion is the assignment of the sensory measurements, in our case object hypothesis H_m, to the currently maintained tracks. Data association is a well discussed problem in the target-tracking literature, e.g. in Hall et al. [7]. Generally the problem scales with exponential complexity and very often heuristics are used to find a promising solution. An optimal solution is the Multiple-Hypotheses-Tracking (MHT) as proposed by Reid [12], which is used in a different context in the proposed architecture.

The key aspect of the MHT-approach is the construction of data-association hypotheses ψ_h^k, that describe the association

of all measurements $H_m(k)$ at time-step k to the current set of track-hypotheses Ω_i^k. Since we operate on a local sensory region per time step and we want to decide upon the relevant objects in this region, we distinguish three cases of association: (a)The data-hypothesis is associated to the currently tracked object of the corresponding object-type, (b)the data-hypothesis is associated to a previously confirmed object or (c)the data-hypothesis is treated as clutter. The object-hypotheses are then used to perform a state-update of the associated track-hypotheses, leading to the new Ω_i^{k+1} in the next time step. In (c) there is no measurement-update performed, only a state-prediction.

In contrast to the original work of Reid [12], we also have to cope with different object categories, therefore one multiple-hypotheses-tracker is maintained for each object-category, which was delivered by the object-recognition system.

For all newly constructed track-hypothesis Ω_i^k their corresponding probability $P_i^k = P(\Omega_i^k, H_m(1:k))$ is computed using Bayes law (see [12])

$$P(\Omega_i^{k-1}, \psi_h^k | H_m(k)) = \frac{1}{c} P(H_m(k) | \Omega_i^{k-1}, \psi_h^k) \cdot P(\psi_h^k | \Omega_i^{k-1}) P(\Omega_i^{k-1} | H_m(1:k-1)) \quad (12)$$

While more and more object-hypotheses are associated, the number of trajectory hypotheses grows exponentially and has to be reduced with different techniques like hypothesis pruning and merging. Furthermore, we limit the total number of hypotheses in order to achieve a more constant computational load.

A track-validation measure is used for deciding that the tracking procedure has converged and the accumulated knowledge about the hypothetical objects is enough for declaring this object as confirmed. This is performed via thresholds of a validation measure, which is defined as

$$M_c = \frac{WPW^T}{c_{obj}^r}, \quad (13)$$

with W as weighting matrix for the state-error covariance matrix P of the EKFs. The value c_{obj} is the average of the last n object-class-confidences delivered by the object-recognition system. The exponent r is chosen from the interval $(0, 1]$ and controls the influence of the classification results.

In figure 2 the time course of a maximum number of 30 hypotheses is shown. The bold (red) lines indicate the best hypothesis in each time-step.

IV. Experiments and Results

In order to show the general feasibility and the performance of the proposed perception architecture, the whole system was implemented using Matlab on a standard PC. The external sensors are a b/w CCD-camera with a resolution of 640×480 Pixel and a horizontal field of view of 30 degree and a LIDAR-sensor with 16 horizontal beams and a field of view of $\pm 15°$ scanning the front-view of the vehicle at a sample-rate of $92ms$. The internal signals used are the ego-velocity, acceleration and the yaw-rate. The processing is synchronized on the slowest source, which is the LIDAR-sensor in the current setup.

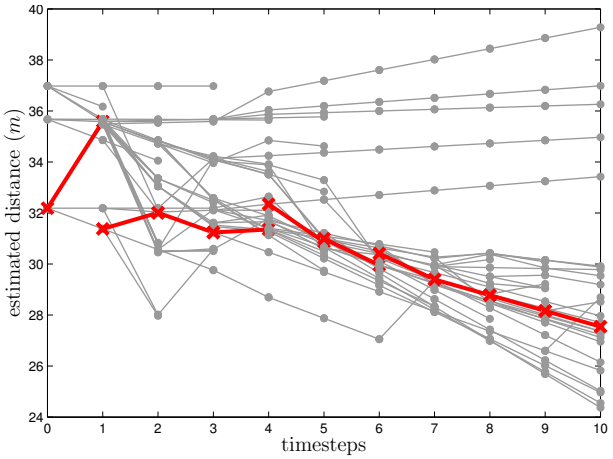

Fig. 2. Time course of the hypothesis. Only the estimated distance is shown (y-axis). The best hypothesis is drawn bold (red).

The system was tested offline with recorded data of different test scenarios. The scenario discussed in this paper deals with a typical construction-site on German Autobahns, where one of the lanes is redirected onto the oncoming roadway, causing an S-shaped path.

The lane width is typically about 3 meters and the signal boards have a distance of 10 meters in the curve and 15-20 meters in front of and behind the curve. This scenario is of moderate complexity and today's ACC-systems and break-assists are usually confused by the low TTC-values to the signal-boards, which can be less than 2 seconds.

Here the task is to detect a stationary car that is immediately located at the end of the S-curve and is invisible when the driver enters the S-curve. The system has to detect and to confirm the vehicle under various initial conditions, so that it should be possible to control a break-assistant based on the confirmed target information delivered by the system. We performed several test-runs on data recorded from a manually prepared construction-site test track that has a left-hand S-shape, see figure 3. The performance of

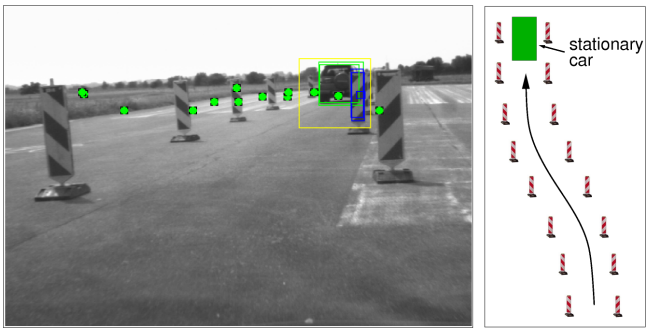

Fig. 3. Left:Vehicle detection using vision and LIDAR, yellow box: current FoA, green box: car hypothesis, blue box: signal-board hypothesis, Right: Plan view of the construction-site with S-shaped roadway

the system was measured in terms of detection distance, confirmation distance and the number of time-steps required for target-confirmation. The detection distance is measured in meter and is the distance to the stationary vehicle at that time when the vehicle was detected by the saliency and object-recognition modules for the very first time. A further distance can be measured at the time of object-confirmation. Since we do not have real ground-truth information about the exact relative position, we used the distance delivered by the LIDAR-sensor as reference (accuracy $\approx 0.2m$).

Three cases of detection and confirmation were distinguished:

1) Detection and tracking using visual information and LIDAR.
2) Detection with LIDAR only, tracking with LIDAR and vision.
3) Detection and tracking with vision only, LIDAR is not used at all.

While case 2 is the usual method in most automotive-perception architectures, case 1 uses the combined sensory saliency processing for detection. Case 3 simulates a break-down of the LIDAR sensor.

The average results of 9 test-runs with ego-speed variations between $30 - 70km/h$ are shown in table II, the numbers after the slash in the table are the standard-deviations of the corresponding measures.

TABLE II

AVERAGE DETECTION AND CONFIRMATION RESULTS

case	**detection distance (m)**	**confirmation distance (m)**	**confirmation-cycles**
vision+LIDAR	35.5 / 3.5	18.8 / 9.2	13.5 / 3.7
LIDAR only	26.3 / 3.8	15.5 / 8.0	11.4 / 3.8
vision only	39.2 / 10.1	18.0 / 6.0	15.1 / 1.8

From table II it is interesting to see that the average detection range of the pure vision processing is in this specific scenario larger than the combined "vision+LIDAR" detection range. This is due to the fact that the combined measurements lead to a broader sensory saliency-distribution and therefore the likelihood for inspecting a specific region decreases when more sensors are used. But if the standard-deviations are taken into account, one can see that the combined detection result is much more reliable than the single sensor detection.

The confirmation distances are between 10 and 15 cycles, which is very close to the minimum of 8-10 cycles that we observed by EKF-simulations on synthetic test-data. The standard deviation of the confirmation distance increases naturally because of the large ego-velocity variations of 40km/h. A more meaningful measure is the number of confirmation-cycles, which is similar in the first 2 cases when the LIDAR-measurements are available. The estimation of the relative velocities based on vision-measurements only is less reliable, leading to an increased number of confirmation-cycles.

It is worth to note that the overall system performance remains at a moderate level in the case of the LIDAR-break-down, when only vision-measurements are available.

V. CONCLUSION

We proposed a hierarchical sensor-fusion concept for the task of environmental perception that addresses the problems of robustness under resource limitations. Robustness was shown against sensor-break-downs and measurement outliers. Further work is required for showing quantified results in terms of detection and false alarm rate. An attention mechanism combined with a multiple hypotheses object recognition and temporal tracking system using different levels of object representations is a possible solution to the robustness-resources dilemma that have to be solved reliably in order to get future ADAS-applications onto the market.

VI. ACKNOWLEDGMENT

The authors would like to thank Sven Bone for preparing the test-drives and Heiko Wersing for providing the object-classification system.

REFERENCES

[1] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision (IJCV)*, 12(1):43–77, 1994.

[2] M. Darms and H. Winner. A modular system architecture for sensor data processing of ADAS applications. In *IEEE Intelligent Vehicles Symposium*, Las Vegas, 6 2005.

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

[4] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer, 2001.

[5] R. Gerber and H.-H. Nagel. 'Occurrence' extraction from image sequences of road traffic scenes. In *Workshop on Cognitive Vision*, ETH Zürich, 9 2002. L. van Gool, B. Schiele.

[6] A. Gern, U. Franke, and P. Levi. Advanced Lane Recognition - Fusing Vision and Radar. In *Proceedings of the IEEE Intelligent Vehicles Symposium*. IEEE, October 2000.

[7] D. L. Hall and J. Llinas. *Handbook of Multisensor Data Fusion*. CRC Press LLC, 2001.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[9] E. Körner, M.-O. Gewaltig, U. Körner, A. Richter, and T.Rodemann. A model of computation in neocortical architecture. *Neural Networks*, 12:989–1005, 1999.

[10] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman and Company, New York, 1982.

[11] K. Naab. Sensorik- und Signalverarbeitungsarchitekturen für Fahrassistenz und Aktive Sicherheit. In *Tagung Fahrerassistenzsysteme: Licht, Sicht und Sicherheit*, Paderborn, 2004. Haus der Technik e.V.

[12] D. B. Reid. An algorithm for tracking multiple targets. *AC*, 24(6):843–854, December 1979.

[13] R. Schweiger, H. Neumann, and W. Ritter. Multiple-cue data fusion with particle filters for vehicle detection in night view automotive applications. In *Proceedings of the IEEE Intelligent Vehicles Symposium*. IEEE, 2005.

[14] T. Strobel and C. Coue. Compendium on sensor data fusion - state-of-the-art of sensors and sensor data fusion for automotive preventive safety applications, 2004. PReVENT Consortium.

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

[16] J. M. Wolfe. Guided search 2.0, a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.