System Approach for Multi-Purpose Representations of Traffic Scene Elements

Jens Schmüdderich, Nils Einecke, Stephan Hasler, Alexander Gepperth, Bram Bolder, Robert Kastner, Mathias Franzius, Sven Rebhan, Benjamin Dittes, Heiko Wersing, Julian Eggert, Jannik Fritsch, Christian Goerick

2010

Preprint:

This is an accepted article published in IEEE 13th Int. Conf. ITSC. The final authenticated version is available online at: https://doi.org/[DOI not available]

System Approach for Multi-Purpose Representations of Traffic Scene Elements

Jens Schmuedderich*, Nils Einecke*, Stephan Hasler*, Alexander Gepperth*, Bram Bolder*,
Robert Kastner†, Mathias Franzius*, Sven Rebhan*, Benjamin Dittes*,
Heiko Wersing*, Julian Eggert*, Jannik Fritsch*, Christian Goerick*

* Honda Research Institute Europe GmbH
Carl-Legien-Str. 30, D-63073 Offenbach, Germany
Email: {firstname.lastname}@honda-ri.de
† Institute for Automatic Control
Darmstadt University of Technology, D-64283 Darmstadt, Germany
Email: {firstname.lastname}@rtr.tu-darmstadt.de

***Abstract*— A major step towards intelligent vehicles lies in the acquisition of an environmental representation of sufficient generality to serve as the basis for a multitude of different assistance-relevant tasks. This acquisition process must reliably cope with the variety of environmental changes inherent to traffic environments. As a step towards this goal, we present our most recent integrated system performing object detection in challenging environments (e.g., inner-city or heavy rain). The system integrates unspecific and vehicle-specific methods for the detection of traffic scene elements, thus creating multiple object hypotheses. Each detection method is modulated by optimized models of typical scene context features which are used to enhance and suppress hypotheses. A multi-object tracking and fusion process is applied to make the produced hypotheses spatially and temporally coherent. In extensive evaluations we show that the presented system successfully analyzes scene elements under diverse conditions, including challenging weather and changing scenarios. We demonstrate that the used generic hypothesis representations allow successful application to a variety of tasks including object detection, movement estimation, and risk assessment by time-to-contact evaluation.**

I. Introduction

A major step towards intelligent vehicles constitutes the research of perception systems whose capabilities equal those of a human driver and which can provide the perceptual basis for the diverse tasks a driver has to fulfill for safe driving.

In this article, we will address the question of how to give a vehicle the ability to achieve reliable perception of the environment even under adverse weather conditions such as rain, night or snow, as well as in a broad spectrum of traffic-scenes such as highway, inner-city and rural roads. We argue that these requirements can only be met by a system approach. More specifically, we will investigate the impact of:

1) Combination of complementary information.
2) Modulation of processing by context information.
3) Generic representations for multiple tasks.

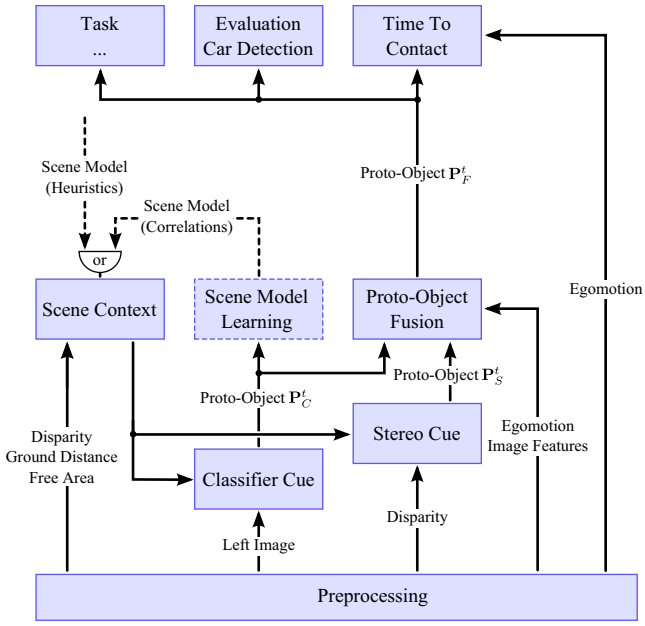
The first point, the combination of complementary sensory information, implies a selection of object specific and unspecific detection methods yielding their peak performance under different environmental conditions or for different object classes. For example stereo pop out can detect arbitrary objects, such as pedestrians, animals, cyclists, or vehicles, as long as they are separable from the depth of their surrounding. In contrast, appearance based detectors can only acquire specific objects but are largely independent from the scene layout.

The second point, the modulation by context information (system-level correlations), addresses the modulation of the detection process or the validation of detection results. For example, the correlation between typical car positions and the position of the road can be used to support car detection in the vicinity of the ground plane.

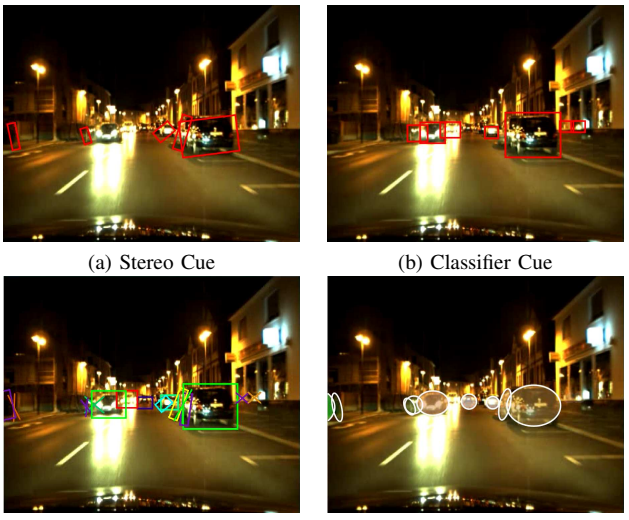
The third point, the representational generality, addresses the question of finding a minimal, i. e. most efficient, representation of scene elements which carries all the information necessary for a variety of system tasks and which affords easy extension to meet the requirements of new tasks.

In the following, we will review the field of related work w. r. t. fulfilling these criteria. A prominent approach for the detection of cars is the implicit shape model based approach by Leibe et al. [1]. Context information in terms of object motion and scene geometry is used, but the use of diverse visual detection cues is not addressed. The system presented by Okutomi et al. [2] shows good performance under broad weather and scene conditions, which is mainly achieved by linking road information to obstacle detection. Szczot et al. [3] also use road information and additional position-size constraints to enhance pedestrian detection. Similar to [4], neither Okutomi nor Szczot approach cue diversity or representational generality.

Strategies for combining visual and non-visual cues have been extensively researched, differing in the information that is fused and the applied fusion method. Commonly, information from different sensors is fused, like sets of



1: System Architecture.



(a) Stereo Cue (b) Classifier Cue

(c) Fused (d) Time To Contact

2: Results of the different processing steps. Red boxes in Figure (a) and (b) represent detections by the respective cues. Rectangles in (c) visualize tracked and fused detections, whereas crosses indicate initial, yet unrelated detections. The color of the ellipses in (d) represents the danger level.

different radar sensors [5]–[7], radar with lidar sensors [5], [8], radar with vision [7], [9], or lidar with vision [10]–[13], including combinations of the above. These approaches gain performance by exploiting the different physical characteristics of different sensors, but the commonly applied sequential processing in which one sensor preselects possible targets for another sensor restricts the modulation capabilities. For pedestrian detection Oliveira et al. [14] accurately evaluate the performance for different sets of visual classifiers, but they neither address context information nor representational generality.

The following sections are structured as follows: In the next section, we will present an overview of a system fulfilling the defined criteria, including a description of its central elements. In Section III we will evaluate the system under a variety of different environmental conditions and show that the chosen design lets the system outperform state-of-the-art object detection systems. To show the generality of the presented approach the system is evaluated exemplary for an object detection task and for time-to-contact analysis.

II. System Architecture and Elements

Figure 1 shows the simplified system architecture. It contains the most important processing elements and will be used to structure this section.

The preprocessing stage (Section II-A) contains the general computation necessary for the different system elements. It comprises the computation of a stereo disparity map, a ground-plane estimation, which is a 3D approximation of the street surface, an ego-motion estimation based on a single track model, and an unmarked street detection, the so called free area.

A central aspect of this system design is the use of object specific and unspecific detection methods. The *Classifier Cue* (Section II-B) is an object specific method selecting regions which are most likely to contain a car. In contrast, the *Stereo Cue* (Section II-C) identifies unspecific obstacles which pop out from the ground-plane. Without loss of generality, the presented system incorporates only visual cues, but targets from other sensors like radar or lidar could be analogously integrated.

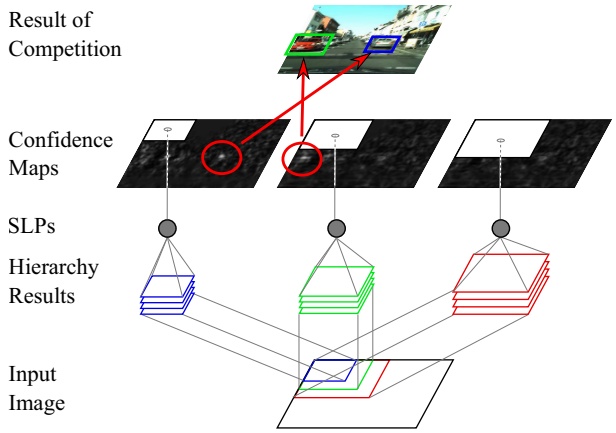
The detection in both the Classifier and the Stereo Cue is modulated by the current scene context using the identified free area, disparity, and distance to the ground plane. This information is used by offline learnt scene context models for early modulation of hypotheses (Section II-D).

To represent the object hypotheses detected by the different cues, the concept of Proto-Objects is chosen. These Proto-Objects can be considered as a pointer to objects in the environment: As long as this pointer remains valid, all visual information can be obtained by referencing this pointer and extracting the information from the image [15].

In the following, the term $P_{C,i}^t$ will be used to refer to the ith Proto-Object of the Proto-Object list $\boldsymbol{P}_C^t$ created by the Classifier Cue at time t, whereas $P_{S,i}^t$ will refer in an analogous way to a Proto-Object created by the Stereo-Cue. In Figure 2a and 2b the detected Proto-Objects by the Stereo- and Classifier-Cue are shown for a typical city night scene.

To keep Proto-Object pointers valid, detections need to be made spatially and temporally coherent. This is the task of the Proto-Object Fusion (Section II-E). It tracks Proto-Objects over time and fuses detections from different cues into a coherent Proto-Object $P_{F,i}^t$. The result of this step is visualized in Figure 2c.

Finally the Proto-Objects are passed to the different tasks, like the detection of cars or a time-to-contact analysis based



3: Car detection by the template-based Classifier. The visual hierarchy is computed for different scales. For each scale and each image-point a SLP classifies the templates as containing a car or a non-car, resulting in a confidence-map. A local competition selects the maxima in these maps.

on the estimated ego-motion and object-motion (Section II-F). Figure 2d shows an exemplary time-to-contact visualization, where color encodes the predicted minimal distance of an object to the egovehicle and saturation encodes the time, when this distance will be reached.
In the following a more detailed description of the different architectural elements is given.

A. Preprocessing

For disparity computation, we use a established stereo implementation, similar to the approach of Fua [16]. It consists of four major steps: In a first step the matching values for all pixels and all disparities are calculated. In the second step the disparity values are interpolated to sub-pixel accuracy by fitting a quadratic curve to the matching values in the neighborhood of the best matching value. The third step is a left-right consistency check for detecting occlusions and the fourth step consists of rejecting small disparity segments.
The approximation of the free area bases on an evaluation of features in a street training region in front of the car and two non-street training regions at the side of the road. By dynamic estimation of probability distributions over these features a pixelwise mask of the unmarked street is obtained. A more detailed description of the approach is presented in [17].
The resulting mask is used to approximate the 3D road surface by a plane. The applied method [18] interprets the input mask as a plane and searches for the optimal plane parameters for matching the mask in the left image to the right image by means of a Hook Jeeves optimization [19].
Based on the obtained 3D representation the distance of each point to the ground-plane is calculated and provided as a pixelwise map.

B. Appearance-based Classifier

The appearance-based classifier generates object hypotheses in three successive steps, which are visualized in Figure 3. First the output of a hierarchical feed-forward architecture as proposed in [20] is computed at multiple scales, resulting in a set of feature maps for each scale and each image-point. Second, for each scale a Single Layer Perceptron (SLP) receives the feature maps around one image point and computes a confidence for this point depicting the center of a car; i. e. if the image-patch used to compute the templates is likely to approximate the boundaries of a car, the confidence value for this point is high, otherwise it is low. In a last step the confidence maps are fed into a competitive selection method that generates a given number of object hypotheses. For this, the local maxima across all scales are detected. Each maximum is directly associated with a Region of Interest (ROI) in the image, where the center of the ROI lies on the maximum and the size of the ROI depends on the scale associated to the corresponding map. The selection works in a greedy fashion. So first, the maximum with the highest confidence is chosen and used to create the Proto-Object $P^t_{C,1}$. With the corresponding ROI the confidence values in all other maps are suppressed. The remaining maxima are then processed in descending order to create the Proto-Objects $P^t_{C,2}$ to $P^t_{C,N}$, whereas each maxima is rejected whose ROI is covered by the already inhibited area by more than 75%. The process stops when the desired number of hypotheses is reached or no further maxima remain. To control the number and quality of detections the confidence maps are thresholded by a value Θ_C. With the choice of the Θ_C a trade-off between false detections and missed cars is made.
This trade-off is supported by incorporating the scene context. It modulates the confidence maps by evaluating the height of a detection above the ground-plane, the maximal overlap with the free area, and by relating the image size of the detection with the 3D size from stereo. The effect of modulation is that local maxima corresponding to implausible hypotheses are ignored.

The SLP is learned in a supervised fashion; Segments containing cars and segments containing non-cars are cropped from the training scenes and normalized in size. The SLP learns to generate high values for positive examples and low values for negative ones. The result is a so called view-tuned-unit [20] which responds robustly to car views of different viewing angle, delivering competitive performance in benchmarks like the UINC car detection [21].

C. Stereo Popout

The instantiation of Proto-Objects by the Stereo Cue constitutes a generic detection method because it does not require object specific knowledge but identifies regions of contiguous depth that stand out from their surroundings. To identify regions of coherent depth, the stereo disparity image is segmented by a region growing approach.

To reduce these detections to relevant objects, each region is related to the current scene context. A depth-region R^t_i

(a) Camera Image (b) Modulation Image

4: A typical modulation image (b) used for incorporating scene context into detection algorithms. Light values in the modulation image relate to plausible locations for vehicles in the image (a), dark values to implausible ones.

which is connected to the current ground-plane, whose 3D height lies in a specific interval, and whose size exceeds a reasonable value, is passed as Proto-Object $P^t_{S,i}$ to the output.

D. Scene Context

Scene context models are incorporated at the level of single object detection methods, e.g. Stereo Cue and Classifier Cue. We have implemented and tested two alternatives: Heuristics and modulation. The major difference between them is their application. Heuristics filter the detection results, whereas modulation can suppress or enhance image regions during the detection process. The underlying assumption of both methods is that for each Proto-Object detection context features can be computed, which are correlated with certain object identities. In other words, specific objects such as cars are characterized by their relation to other scene-elements or processing results. For example, cars are characterized by a high proximity to the ground-plane or a specific physical size.

In the case of heuristics, the representation of context models describing such characteristics boils down to the learning of appropriate thresholds. Optimizing these thresholds is performed on a small training set derived from parts of the overcast, sunny and rain stream (more details on these streams follows in Section III). Unfortunately, the learning of the thresholds constitutes a multi-objective optimization problem because the single thresholds depend on each other. For the three features used for the Stereo Cue it is possible to find the optimal thresholds by a brute force search. However, for the Classifier Cue we used six different features. In this case, we applied a standard evolutionary optimization algorithm [22] to find a good solution in reasonable time.

The downside of heuristics is that they constitute a binary decision. Hence, this approach may fail if the characteristics of the scene, and by this also the optimal thresholds, change too much. This problem is addressed by the context modulation. In contrast to the heuristics the scene context models used for modulation reflect the statistical dependencies between context features and objects. These dependencies are learned autonomously using methods described in [23]. In a conceptually new step, learned dependencies are inverted to produce hypothesis feature distributions given the object type "vehicles". Such distributions can be transformed to *modulation images* (see Figure 4) which are multiplied with the confidence maps produced by the classifier (see Section II-B) to achieve selection of vehicle-containing image regions.

The achieved improvement by means of context features is discussed in Section III.

E. Proto-Object Fusion

The Proto-Objects $\boldsymbol{P}^{t_c}_C$ and $\boldsymbol{P}^{t_s}_S$ represent instantaneous, asynchronous detections, independent from each other and the temporal history. The Proto-Object Fusion relates these detections over time and space, thus building coherent representations. The approach comprises two major steps: First, the tracking of detections in the isolated cues and second the fusion of detections in different cues relating to the same physical object.

The aim of the tracking in isolated cues is relating all Proto-Object detections $\{P^{t_0}_i, P^{t_1}_i, \ldots, P^{t_N}_i\}$ to one Proto-Object model $\hat{P}^t_i$ [1]. Therefore the Kalman-Filter based multi-object tracking presented in [15] is extended to cope with the high dynamics of the automotive domain. In the traditional approach Proto-Object $\hat{P}^t_k$ is associated with a detection $P^{t_c}_i$ if

$$k = \underset{j}{\operatorname{argmax}} \left(S(P^{t_c}_i, \hat{P}^{t_c}_j) \right) \quad (1)$$

$$\wedge \quad S(P^{t_c}_i, \hat{P}^{t_c}_k) > \Theta_F \quad (2)$$

where $S(P_i, P_k) \in [0, 1]$ is a similarity function between the Proto-Objects P_i and P_k, and $\hat{P}^{t_c}_k$ is a prediction of the Proto-Object $\hat{P}^t_k$ towards the current timestep t_c. Thus, the decision to bind an existing model to a detection is made by predicting all available models towards the timestep t_c of the detection and perform a comparison between the detection and the predicted models. If a detection is associated to a model, the model is updated as described in [15].

This traditional approach exhibits three limitations for the automotive domain:

1) Binding models to predictions is a one-to-one mapping.
2) S depends on the overlap in the image only.
3) S does not incorporate dynamic uncertainties.

The approach presented in this paper binds detections to the best fitting model, thus overcoming the one-to-one mapping. As the similarity measure based on visual overlap can neither cope with occlusions by other objects, nor with the noisy cue segmentation, we refine the similarity function between two Proto-Objects P_i and P_k as

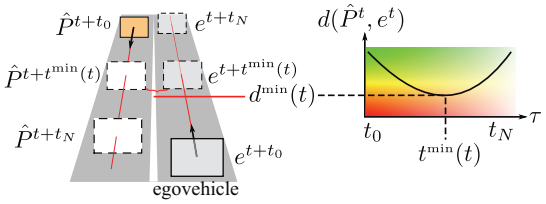
$$\begin{aligned} S(P_i, P_k) = w_P \cdot S_P(P_i, P_k) + w_O \cdot S_O(P_i, P_k) \\ + w_C \cdot S_C(P_i, P_k) + w_S \cdot S_S(P_i, P_k) \ , \end{aligned}$$

where $S_P(\cdot, \cdot)$ encodes the similarity based on 3D position, $S_O(\cdot, \cdot)$ based on 2D overlap, $S_C(\cdot, \cdot)$ based on color, and $S_S(\cdot, \cdot)$ based on Local Orientation Coding (LOC) features describing the image structure [24]. The w_i are fixed scalar weightings of the respective similarity measures, summing

[1] As the tracking is identical for each Cue, the source-indices have been dropped for convenience

(a) overcast day (b) sunny day (c) night (d) rain (e) snow

5: These images are example frames of the five different weather conditions we used for evaluating our system.



6: Visualization of Time To Contact estimation. The smallest relative distance $d_i^{\min}(t)$ and the time $t_i^{\min}(t)$ when this distance is reached are evaluated and determine the color coding. In this figure the indices are dropped for convenience.

up to 1.0. The similarity measures $S_P(\cdot,\cdot)$ and $S_O(\cdot,\cdot)$ incorporate the dynamic process- and measurement-covariances of the Kalman-Filter, which are chosen such, that they account for uncertainties based on an object's distance, the ego-motion, and the temporal history.

As a result of this step, the detections of the Classifier- and the Stereo-Cue are aggregated to Proto-Objects $\hat{P}_{C,i}^t$ and $\hat{P}_{S,i}^t$ respectively. In a second step, these Proto-Objects are fused into one list of coherent Proto-Objects $\boldsymbol{P}_F^t$. Here the fusion is done exactly as in the above step, except that the similarity is not computed between detections $P_{C,i}^t$ and and tracked Proto-Objects $\hat{P}_{C,i}^t$, but rather between tracked Proto-Objects $\hat{P}_{C,i}^t$, $\hat{P}_{S,i}^t$ and fused Proto-Objects $P_{F,i}^t$.

F. Time To Contact Evaluation

The time to contact evaluation demonstrates the applicability of the perceived Proto-Objects to various tasks, including risk assessment. This risk is composed of two estimations: The minimal predicted distance $d_i^{\min}(t)$ between the egovehicle and an object $P_{F,i}^t$ in a time window $[t+t_0, t+t_N]$, and the time $t_i^{\min}(t)$ when this minimal distance will be reached. They are estimated by

$$d_i^{\min}(t) = \min_{\tau\in[t_0,t_N]} d(\hat{P}_{F,i}^{t+\tau}, e^{t+\tau}) \tag{3}$$

$$t_i^{\min}(t) = \operatorname*{argmin}_{\tau\in[t_0,t_N]} d(\hat{P}_{F,i}^{t+\tau}, e^{t+\tau}) \ , \tag{4}$$

where $e^{t+\tau}$ is the state of the egovehicle predicted for time $t+\tau$, $\hat{P}_{F,i}^{t+\tau}$ is a Proto-Object predicted towards time $t+\tau$, and $d(\hat{P}_{F,i}^t, e^t)$ is the distance between the closest points of $\hat{P}_{F,i}^t$ and the egovehicle at time t. The predictions incorporate current position and velocity gained from the Kalman Filter [25].

Figure 2d shows a visualization of this risk estimation by color coding. Here the minimal predicted distance is mapped to hue, and the minimal time to saturation. This coloring scheme is also visualized in Figure 6.

III. Experiments

In order to assess the performance of our system, we set up two different tasks. The first task is to detect all cars within the current scene at a low false positive rate. With this experiments we evaluate two aspects:

1) Is there a gain in using unspecific detection cues when searching for specific objects?
2) Is there a quantitative gain in using context modulation?

We evaluate robustness against different weather conditions (overcast day, sunny day, night, rain, snow (see Figure 5)) as well as generality over different scene types (inner city, highway, rural road, industrial area). To evaluate our system w.r.t. these requirements, we recorded five video streams of roughly 10 minutes length using an experimental prototype vehicle. For all streams the same route was traversed, encompassing all of the aforementioned scene types and weather conditions. Hence, these streams allow a sophisticated analysis of the raised questions.

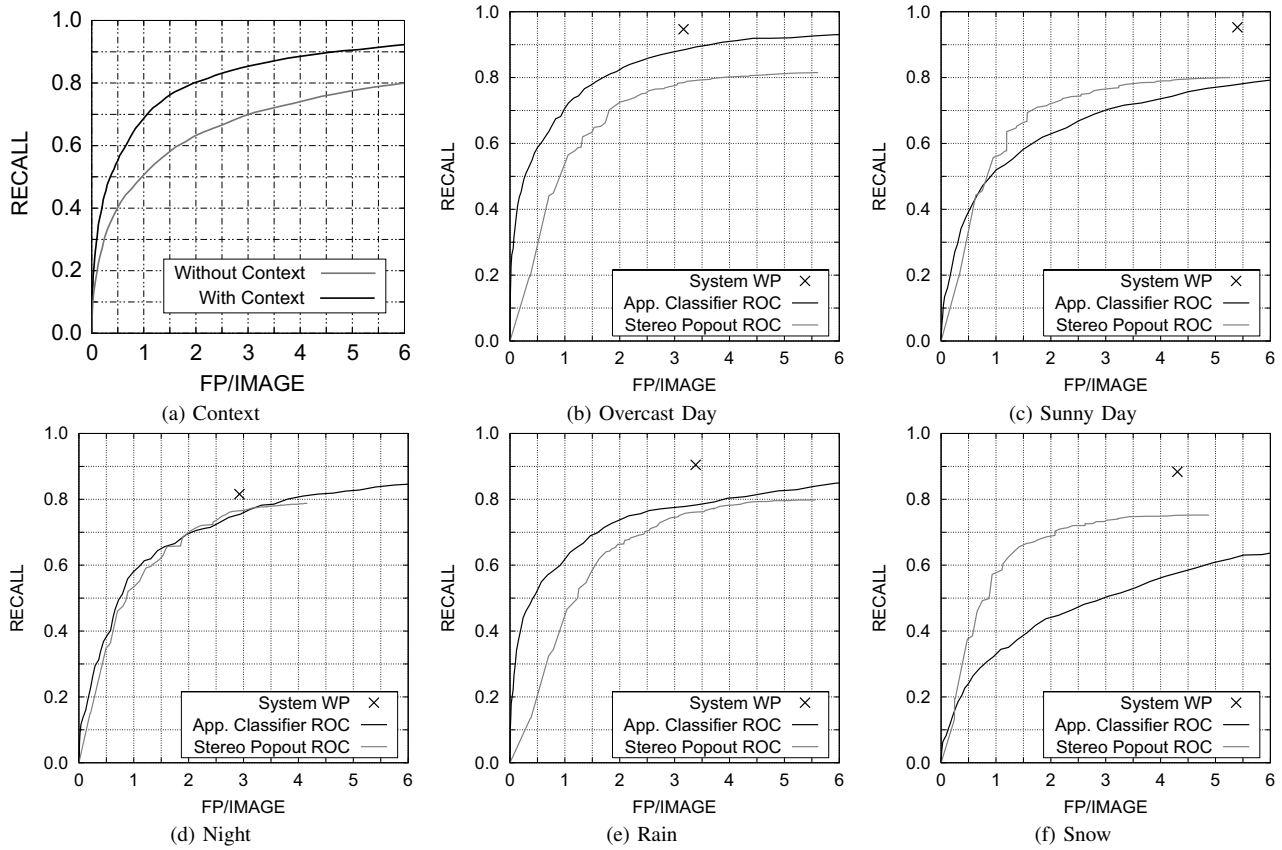
Parts of the overcast, sunny and rain stream are used to train the learning modules of the system. The remaining parts of these streams as well as the full night and snow streams are used for evaluation.

In a second experiment, the system has to analyze the current scene and assign a risk to each object, based on the previously introduced time to contact analysis. This task requires a certain amount of scene understanding as objects have to be identified and their movement has to be predicted.

A. Car Detection Evaluation

As mentioned above, our evaluation of car detection is done on five different streams. For quantitative evaluation, one image per second has been annotated by hand to create ground truth data. The annotation constitutes a rectangular mask for each car in the scene, approximating the shape of each car. These masks are *intrinsic*, that is, they approximate the actual object shape even if the object is not completely visible. Additionally, the amount of occlusion for each object is labeled.

The performance of single methods for the car detection task is assessed through Receiver Operator Characteristic (ROC) analysis. To investigate the trade-off between correct detections and false positive detections of non-car scene elements, we evaluate the false-positive rate per image frame



(a) Context (b) Overcast Day (c) Sunny Day

(d) Night (e) Rain (f) Snow

7: a) ROC comparison of the average (over all streams) Classifier Cue performance with and without using scene context models b-f) Comparison of selected Classifier Cue and Stereo Cue performance with System Working Point (WP) performance for each stream.

(FP/IMAGE) with respect to the so called RECALL R, which is defined as

$$R = \frac{TP}{TP + FN} \quad . \tag{5}$$

Here, TP stands for true-positives, the number of labeled cars which were detected by the system, and FN stands for false-negatives, the number of labeled cars which were not detected by the system. A labeled car is considered detected if the center of one of the generated Proto-Objects lies inside the annotated region. The advantage of using ROCs lies in analyzing the detection methods across their whole working range instead of using a single working point. Unfortunately, this is not possible for the whole system because contrasting to the classifier, the system's performance does not depend on one parameter but on a large set of parameters. Thus the creation of a ROC would require a complete iteration of all combinations of parameters. Hence, we run the system with optimal parameters derived from isolated ROC analyses of the single cues.

We performed a thorough analysis of single cue performance[2] and compared the results to the performance of the whole system. When comparing two ROC curves, the one ROC curve being above the other is considered to be better because it shows continuously better performance. In case a working point is compared to an ROC curve, we consider the performance of the working point to be better if it is above the ROC curve because it constitutes a better trade-off between false-positive rate and recall. As a complete review of our results would go beyond the scope of this paper, we restrict the results to the most striking ones.

First, the results show that inclusion of scene context models dramatically improves the performance of the individual cues as shown here for the Classifier Cue. The car classifier employed in our system is appearance-based, i.e. its response depends on local image information. Due to this, the classifier itself is not able to revoke implausible detections according to their position or size. Indeed, a street scene has a defined structure and allows for various context information like cars being close to the ground plane, cars having a certain physical size or cars occurring only in certain areas of an image. Figure 7a shows that such scene context models significantly boost the performance of the car classifier.

The second conclusion from the results is that even for the detection of vehicles the complete system has a significantly higher performance than the appearance based classifier alone. As mentioned above, it is unfeasible to create a ROC of the whole system because of the

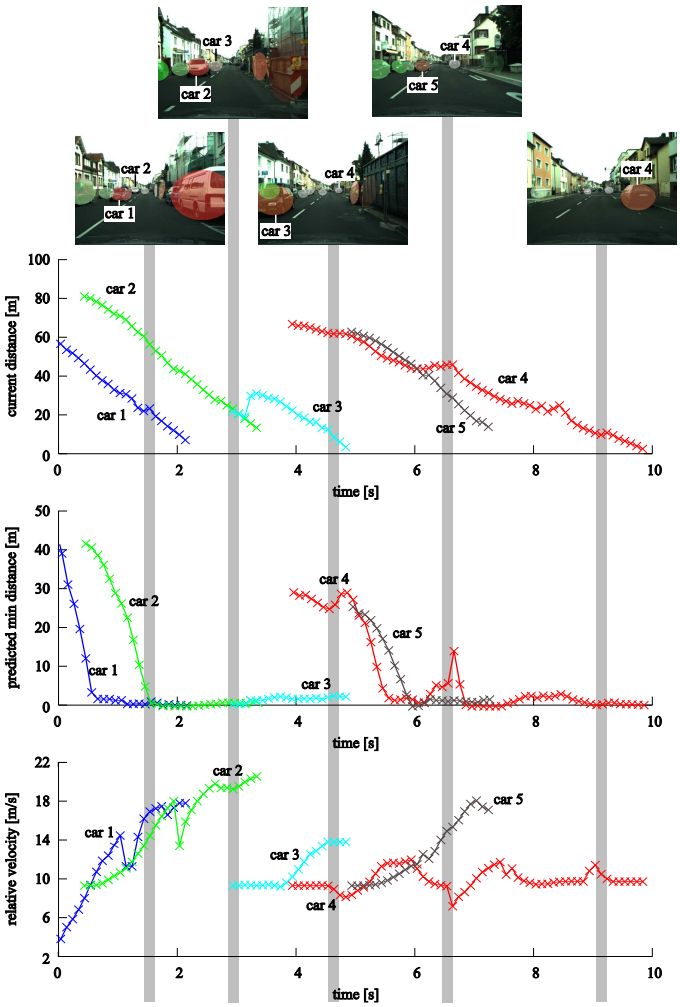
[2]Without scene context for the Classifier Cue, with scene context for the Stereo Cue.

large parameter space and the time required for a single performance evaluation. For this reason, we compare the ROCs of the single cues with the working point of the whole system. The results are displayed in the Figures 7b, 7c, 7d, 7e, and 7f for the different weather conditions. All these plots show a significantly higher performance of the whole system compared to the single cues. On average, the recall of the whole system is 0.1 higher than the individual cues, at equal false positive rate. However, this effect varies considerably for the different weather conditions. The highest gain is achieved for the sunny day stream in Figure 7c, where the system achieves a recall of 0.95 compared to a recall of 0.8 for the classifier. The lowest gain in recall rate appears at night, as visualized in Figure 7d. Here, the whole system has a recall of 0.81 at a recall of 0.77 for the best single cue. These results indicate that methods for dynamical fusion, dependent on the weather conditions and cue performance require further investigation.

Two important conclusions can be derived from this experiment: First, the use of scene context greatly improves detection performance. Second, even when searching for specific objects, the use of unspecific detection cues improves the performance significantly. However, a dynamic fusion strategy incorporating knowledge about cue performance under different weather conditions is suggested. This has also been verified in first tentative experiments.

B. Time To Contact Evaluation

A quantitative analysis of the risk assessment is not possible as no ground-truth data exists and the use of simulators implies the problem of insufficiently reproducing noise and error sources of real-world environments. For this reason, we present a qualitative evaluation here. The results are plotted for a 10 second interval in an exemplary inner-city day-scene in Figure 8. Car 1 is initially detected at a distance of 56m. The initial relative velocity in z-direction (longitudinal movement) is estimated to be 4m/s, thus the minimal predicted distance within 4 seconds is 40m. Our Kalman Filter approach for tracking hypotheses (see Section II) requires approximately 12-20 frames until convergence, which is expressed in a slowly increasing relative velocity. With increasing relative velocity, the minimal predicted distance drops to below 1m but not below 0m. This is the effect of evaluating the distance in a 3D coordinate system. Car 2 is already detected at a distance of 80m, giving the Kalman Filter enough time to converge, so that a minimal distance below 1m can already be predicted at a distance of 56m. At about $t = 2$ the velocities change abruptly. Here, the car is lost by the Classifier Cue, and the Stereo Cue takes over. As the Stereo Cue usually detects objects later than the Classifier Cue, the estimated velocity is lower. However, the system recovers from this loss after about 4 frames. Car 3 following car 2 at a close distance is detected late. Thus, its prediction becomes accurate at a distance of only 12m. In contrast to the other cars, the relative velocity of the parked car 4 is only caused by the egovehicle's movement. Thus



8: Results of the time to contact analysis over a 10s inner-city day-sequence. The images contain the risk colored as described in the previous section. For selected objects, this plot shows the currently measured distance (top row), the minimal predicted distance estimated relative velocity in z-direction (bottom row).

the predicted minimal distance falls off much slower. Since all the cars on the opposing lane move out of the camera's field of view before they reach the egovehicle, the measured current distance never falls below 3m.

This experiment shows the feasibility of the presented system for risk assessment. The settling time of the Kalman Filter between 12-20 frames is slightly higher than reported by [26]. We assume, that this is caused by a less accurate depth, the missing use of optical flow for initialization, and an unconstrained 3D coordinate space which is more error-prone than the commonly used 2D-projections to the ground plane. As reported in [26], we believe that the settling time can additionally be improved by incorporating velocity information from the radar-sensor.

IV. Conclusion and Outlook

In this paper we presented a system whose design differs in three major aspects from state-of-the-art approaches: It

includes several complementary visual object detection methods (cues) generating vehicle hypotheses that are tracked and fused to obtain coherent representations of objects in the environment. Moreover, the cues are modulated by incorporating context information, like the relation of detections to the ground plane or to the estimated drivable area. Finally, the concept of Proto-Objects is employed, serving as a generic, common representation which affords the applicability to various tasks.

In extensive evaluations we compared our proposed system with a state of the art appearance based classifier. These experiments demonstrate that the combination of object specific and unspecific detection cues is beneficial, even for the detection of specific objects: The presented system significantly outperforms the appearance based classifier under various conditions.
Moreover, the use of car-unspecific detection cues equip the system with the ability to detect unexpected obstacles, such as pedestrians, cyclists, motor-bikes or animals.
It is worth noting that the gain of incorporating multiple cues varied among the different scenes, indicating a not yet sufficient fusion strategy.

In an exemplary experiment we demonstrated the feasibility of the chosen representations for a typical risk assessment based on trajectory prediction. The shown performance is convincing, considering that the system has not been specially designed for this task, but only evaluates the obtained Proto-Object representations.

Similar to previous approaches we could show that incorporating context information to the cue detection significantly improves performance. However, in the current approach this is mainly exploited for the car detection, but much less for the remaining processing, like tracking and fusion. Methods for improving these processes by incorporating such context information or top-down knowledge are subject to further research. They would constitute an important step towards scene understanding as they would enable the system to go beyond visual similarities and maintain coherent representations even if objects are temporarily occluded.

V. Acknowledgements

We would like to thank our colleagues from Honda R&D, especially Marcus Kleinehagenbrock for the valuable comments and discussions and for making the experiments possible.

References

[1] B. Leibe, N. Cornelis, K. Cornelis, and L. J. V. Gool, "Dynamic 3D Scene Analysis from a Moving Vehicle," in *CVPR*, 2007.

[2] A. Seki and M. Okutomi, "Robust obstacle detection in general road environment based on road extraction and pose estimation," in *Proc. of the IEEE Intelligent Vehicles Symposium*, 2006, pp. 437–444.

[3] M. Szczot, O. Lohlein, M. Serfling, and G. Palm, "Incorporating contextual information in pedestrian recognition," in *Proc. of the IEEE Symposium on Intelligent Vehicles*, 2009.

[4] T. N. Nguyen, M. M. Meinecke, M. Tornow, and B. Michaelis, "Optimized grid-based environment perception in advanced driver assistance systems," in *Proc. IEEE Intelligent Vehicles Symposium*, 3–5 June 2009, pp. 425–430.

[5] M. S. Darms, P. E. Rybski, C. Baker, and C. Urmson, "Obstacle Detection and Tracking for the Urban Challenge," *IEEE Intelligent Transportation Systems*, vol. 10, no. 3, pp. 475–485, Sept. 2009.

[6] J. Dickmann, F. Diewald, M. Maehlisch, J. Klappstein, S. Zuther, S. Pietzsch, and S. Hahn, "Environmental Perception For Future Integrated Safety Systems," in *Proc. 21st Int. Tec. Conf. Enhanced Safety of Vehicles*, June 2009.

[7] R. Schubert, G. Wanielik, and K. Schulze, "An analysis of synergy effects in an omnidirectional modular perception system," in *Proc. IEEE Intelligent Vehicles Symposium*, 3–5 June 2009, pp. 54–59.

[8] M. Skutek, "Ein PreCrash-System auf Basis multisensorieller Umgebungserfassung," Ph.D. dissertation, TU Chemnitz, 2006.

[9] T. Giebel, M. M. Meinecke, M. A. Obojski, M. Gonter, and U. Widmann, "Current Trends in Vehicle Active Safety and Driver Assistance Development," in *In Proc. FISITA*. Springer Automotive Media, 2008.

[10] L. Huang and M. Barth, "Tightly-coupled LIDAR and computer vision integration for vehicle detection." in *Proc. IEEE Intelligent Vehicles Symposium*, 3–5 June 2009, pp. 604–609.

[11] N. Kaempchen, K. C. Fuerstenberg, A. G. Skibicki, and K. C. J. Dietmayer, "Sensor fusion for multiple automotive active safety and comfort applications," *Advanced Microsystems for Automotive Applications*, pp. 137–163, 2004.

[12] M. Mahlisch, R. Schweiger, W. Ritter, and K. Dietmayer, "Sensorfusion Using Spatio-Temporal Aligned Video and Lidar for Improved Vehicle Detection," in *Proc. IEEE Intelligent Vehicles Symposium*, 2006, pp. 424–429.

[13] H. Weigel, P. Lindner, and G. Wanielik, "Vehicle tracking with lane assignment by camera and lidar sensor fusion," in *Proc. IEEE Intelligent Vehicles Symposium*, 3–5 June 2009, pp. 513–520.

[14] L. Oliveira, U. Nunes, and P. Peixoto, "On Exploration of Classifier Ensemble Synergism in Pedestrian Detection," *IEEE Intelligent Transportation Systems*, vol. 11, pp. 16–27, 2010.

[15] J. Schmuedderich, H. Brandl, B. Bolder, M. Heracles, H. Janssen, I. Mikhailova, and C. Goerick, "Organizing Multimodal Perception for Autonomous Learning and Interactive Systems," in *IEEE-RAS International Conference on Humanoid Robots*, Daejeon, Korea, December 2008, pp. 312–319.

[16] P. Fua, "A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features," *Machine Vision and Applications*, vol. 6, no. 1, pp. 35–49, 1993.

[17] 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.

[18] N. Einecke, S. Rebhan, V. Willert, and J. Eggert, "Direct Surface Fitting," in *International Conference on Computer Vision Theory and Applications, in press*, 2010.

[19] R. Hooke and T. A. Jeeves, ""Direct Search" Solution of Numerical and Statistical Problems," *Journal of the Association for Computing Machinery*, vol. 8, no. 2, pp. 212–229, 1961.

[20] 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.

[21] H. Wersing, S. Kirstein, B. Schneiders, U. Bauer-Wersing, and E. Körner, "Online Learning for Bootstrapping of Object Recognition and Localization in a Biologically Motivated Architecture." in *IEEE International Conference on Computer Vision Systems*, 2008, pp. 383–392.

[22] C. Igel, T. Suttorp, and N. Hansen, "Steady-state Selection and Efficient Covariance Matrix Update in the Multi-objective CMA-ES," in *Proceedings of the Fourth International Conference on Evolutionary Multi-Criterion Optimization*, 2007, pp. 171–185.

[23] A. Gepperth, J. Fritsch, and C. Goerick, "Cross-module learning as the first step towards a cognitive system concept," in *Processings of the Second International Conference on Cognitive Systems*, 2008.

[24] C. Goerick, D. Noll, and M. Werner, "Artificial Neural Networks in Real Time Car Detection and Tracking Applications," *Pattern Recognition Letters*, vol. Vol. 17, pp. pp.335–343, 1996.

[25] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge: MIT Press, 2005.

[26] A. Barth and U. Franke, "Estimating the Driving State of Oncoming Vehicles From a Moving Platform Using Stereo Vision," *IEEE Intelligent Transportation Systems*, vol. 10, no. 4, pp. 560–571, Dec. 2009.