Proceedings of the 16th International IEEE Annual Conference on
Intelligent Transportation Systems (ITSC 2013), The Hague, The
Netherlands, October 6-9, 2013

Efficient Scene Understanding for Intelligent Vehicles Using a Part-Based Road Representation

Daniel Töpfer[1], Jens Spehr[1], Jan Effertz[1] and Christoph Stiller[2]

Abstract— **In this paper we propose a novel part-based approach to scene understanding, that allows us to infer the properties of traffic scenes, such as location and geometry of lanes and roads. Lanes and roads are parts of our undirected graphical model in which nodes represent parts or sub-parts of scenes and edges represent spatial constraints. Spatial constraints are statistically formulated and allow us to take advantage of low-level relations as well as high-level contextual information. The estimation of scene properties is formulated as an inference problem, which is solved using non-parametric belief propagation. Inferring about high-level scene properties, while relying on error-prone sensory cues is challenging and computational expensive. Therefore, we introduced a novel depth-first message passing scheme. This scheme is applied to several challenging real world scenarios showing robust results and real-time performance.**

I. INTRODUCTION

The ability of sensing and understanding the vehicles environment is a key technology for autonomous driving and Advanced Driver Assistance Systems (ADAS). Each of these applications require a robust estimation of geometrical and topological scene properties like e.g. location, course and number of lanes in the vehicles environment as well as topological relations. Retrieving such high-level information from sensory cues is extremely difficult due to their error-proneness and ambiguities.

Many scene understanding approaches rely on one specific computer vision approach or one sensory cue [1]. Applied to real world applications, these approaches are expected to have a poor performance. To overcome this issue, recent approaches make use of a combination of different sensory cues and low-level vision approaches [2], [3], [4]. Similarly, our approach allows us to use different low and high-level sensory cues.

Another important requirement is the ability to model uncertainties as well as contextual, spatial and semantic relations between objects. The drawback of recent approaches is that they only consider high-level contextual relations like e.g., ground, sky, and walls [5] and thus they can not benefit from low-level relations like e.g., the spatial configuration of local vision features. However, we found that by incorporating low-level relations we can greatly improve the robustness of our approach.

II. RELATED WORK

Many approaches treat scene understanding as a segmentation problem. Bileschi [6] proposed a method, which segments street scenes in classes, such as cars, pedestrians, roads and trees using an biologically inspired image representation. A conditional random field is proposed by Wojek et al. [7] to jointly perform object detection and scene labeling. Sturgess et al. [4] developed a segmentation of road scenes based on appearance cues and structure-from-motion features. Another segmentation approach is presented in Ess et al. [8]. Their traffic scene segmentation allows them to assign semantic labels like road types, cars or pedestrian crossings to individual segments.

Existing approaches for high-level scene understanding often use generative graphical models. Wang et al. [9] proposed a hierarchical Bayesian network to perform activity detection in traffic scenes from a static platform. A dependent Dirichlet processes is used in [10] to understand the behavior of moving object in scene. A generative model for 3d scene interpretation was proposed by Wojek et al. [3]. Their model jointly performs multi-class object detection, object tracking, scene labeling and 3d geometric relations. For inferring about 3d scene context as well as 3d multi-objects a reversible-jump Markov Chain Monte Carlo (MCMC) scheme is employed. Geiger et al. [2] also proposed the use of reversible-jump MCMC to infer geometrical, topological properties of scenes as well as semantic activities.

Closely related to our approach is the work of Spehr et al. [11], where a part-based scene understanding approach for parking lots is proposed. They proposed a hierarchical decomposition of a parking-lot scene into geometrical primitives like u-shapes and l-shapes. These primitives are again decomposed into simple features like lines.

Similarly, we propose a hierarchical model for inferring about more general traffic scenes. In our approach complex traffic scenes are decomposed into roads and lanes. Lanes are again decomposed into patches representing e.g. two parallel lane-marking features. The lowest level of our model is representing observable sensory cues like e.g. road-edges or lane-markings. These low-level sensory cues incorporate noise and clutter into our model, which makes performing inference challenging. Towards this goal we developed a novel depth-first message passing scheme, which allows us to perform inference over complex scenes in real-time.

III. PART-BASED SCENE UNDERSTANDING

In this section we present our part-based approach for scene understanding. We start by brief introducing our proba-

[1] D. Töpfer, J. Spehr., J. Effertz are with the Driver Assistance and Integrated Safety Department, Group Research, Volkswagen AG, D-38436 Wolfsburg, Germany [daniel.toepfer, jens.spehr, jan.effertz]@volkswagen.de

[2] C. Stiller is with the Department of Measurement and Control, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany stiller@kit.edu

978-1-4799-2914-613/$31.00 ©2013 IEEE

bilistic representation of lanes, roads and traffic scenes. In the following we present our approach to performing inference including non-parametric belief propagation, belief sharing and our novel depth-first message passing scheme.

A. PART-BASED ROAD MODEL

Our approach to scene understanding is based on an undirected graphical model (see Fig. 3). Generally, graphical models capture the way, joint distributions over random variables can be decomposed into a product of factors. Each of these factors only depends on a subset of the variables. This local decomposition leads to efficient inference algorithms. As can be seen in Fig. 3 our graphical model is composed of a set of hidden $\mathbf{x} = \{\mathbf{x}_1 \ldots \mathbf{x}_N\}$ and observable $\mathbf{y} = \{\mathbf{y}_1 \ldots \mathbf{y}_N\}$ random variables. Hidden random variables represent complex objects (e.g.. lane, roads) or parts of objects. In our model, all parts and objects have a continuous and multidimensional state vector, which defines the position and orientation of objects in two-dimensional Euclidean space.

Edges in our graphical model represent the relationship between pairs of random variables. The relationship between two hidden variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is encoded by an associated pairwise potential $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$, representing their spatial compatibility. Furthermore, we introduced observation potentials $\phi_i(\mathbf{x}_i, \mathbf{y}_i)$ encoding the relationship between a hidden variable $\mathbf{x}_i$ and an observable variable $\mathbf{y}_i$.

Formally, our graphical model is defined by the graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ with nodes $\mathcal{V}$ and edges $\mathcal{E}$. Nodes represent our model parts and edges $\mathcal{E} = \{\mathcal{E}_\psi, \mathcal{E}_\phi\}$ observation potentials and spatial constraints. Accordingly, the joint probability distribution over all random variables in our graphical model can be written as

$$\begin{aligned} &-\log\left(p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_p}|\mathbf{y})\right) \\ &= \log(Z) + \sum_{(i)\in\mathcal{V}} \Phi_i(\mathbf{x}_i, \mathbf{y}_i) + \sum_{(i,j)\in\mathcal{E}_\psi} \Psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \quad (1)$$

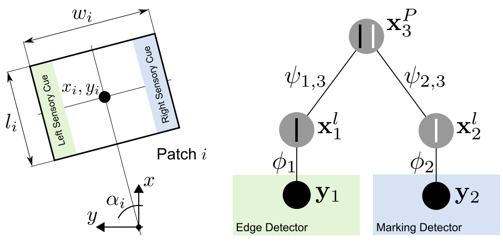
where $\Phi_i = -\log(\phi_i)$, $\Psi_{ij} = -\log(\psi_{ij})$ and the real number Z denotes the partition function that normalizes the probability distribution.

1) REPRESENTATION OF LANES AND ROADS:

The basic parts of our graphical model are local patches representing finite areas of road scenes (see Fig. 2). As can be seen in Fig. 1a patches are defined by a left and a right sensory cue corresponding to lane-markings or road-edges (e.g. curbstones, crash barriers, greensward etc.). This gives us the opportunity to apply patches to a variety of scenes (e.g. highway, rural or urban roads).

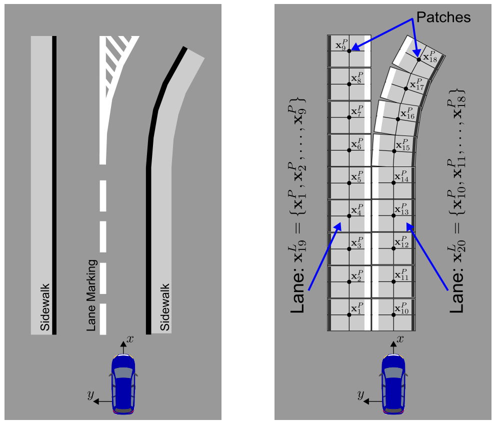
Formally, patches are defined by a five-dimensional state vector $\mathbf{x}_i^P = (x_i, y_i, \alpha_i, w_i, l_i)$. The parameters x_i, y_i and α_i define the configuration of a patch in a local vehicle centred coordinate frame, where $(x_i, y_i) \in R^2$ is the patches position and $\alpha_i \in [0, 2\pi)$ its orientation angle. Additionally, l_i and w_i define the length and the width of a patch.

Another important object of our part-based model are lanes (see Fig. 2). In order to describe lanes, often specific geometrical representations are selected such as clothoid for highways or splines for intersections [2].



(a) Parametrization (b) Part-based model

Fig. 1: Parametrization and part-based model of a patch


Sidewalk
Lane Marking
Sidewalk
x
y
Patches
Lane: $\mathbf{x}_{19}^L = \{\mathbf{x}_1^P, \mathbf{x}_2^P, \ldots, \mathbf{x}_9^P\}$
Lane: $\mathbf{x}_{20}^L = \{\mathbf{x}_{10}^P, \mathbf{x}_{11}^P, \ldots, \mathbf{x}_{18}^P\}$
x
y

(a) (b)

Fig. 2: Example for a patch-based lane representation (b) of a road with two lanes (a).

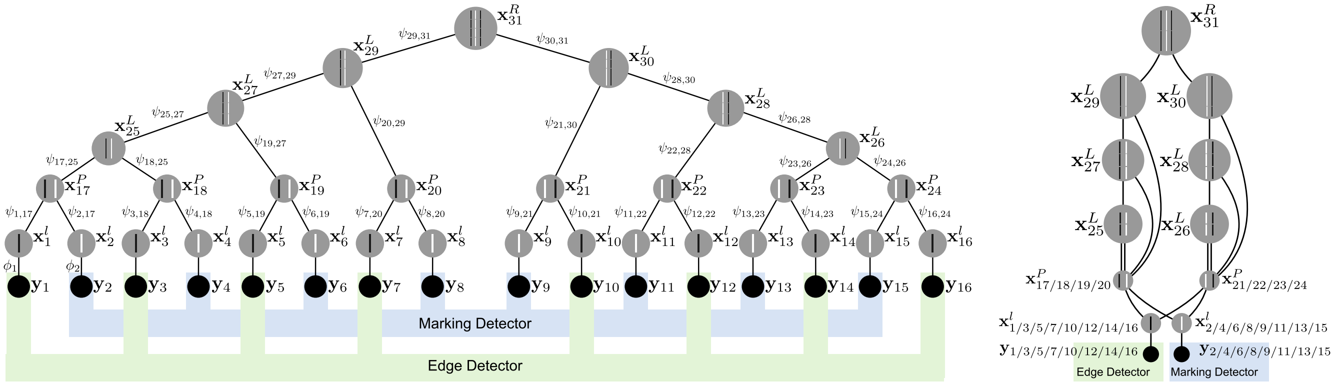
However, in our approach, a major requirement is that our lane representation is applicable to a variety of different scenarios. Furthermore, generating lanes from their sub-parts has to be computational inexpensive. Therefore, we propose to represent lanes by a set of n_P individual patches given by $\mathbf{x}^L = \{\mathbf{x}_1^P, \mathbf{x}_2^P, \ldots, \mathbf{x}_{n_P}^P\}$ (see Fig. 2b). Besides the ability of describing various lane geometries, this representation avoids to introduce additional object parameters, which minimize the computational complexity during inference. However, some application may require a smooth lane representation. In that case, we can easily adopt our model by introducing additional object parameters (e.g. curvature or curvature rate).

As shown in Fig. 3a, lanes are once again used as parts of more complex objects representing road scenes. Since we apply our model to very different scenes, such as multi-lane roads, highway exit-ramps or urban roads, a flexible scene representation is required. Generally, road scenes are defined by a set of n_l lanes and are given by $\mathbf{x}^R = \{\mathbf{x}_1^L, \mathbf{x}_2^L, \ldots, \mathbf{x}_{n_l}^L\}$. In order to distinguish between scenes with e.g. parallel, splitting or merging lanes we introduce dedicated spatial constrains. An example is given in Fig. 3a, where the potentials $\psi_{29,31}$ and $\psi_{28,31}$ ensure a parallel lane configuration.

2) SPATIAL CONSTRAINTS:

One key aspect to model lanes and roads as a graphical model is the formulation of local spatial constraints between pairs of hidden random

978-1-4799-2914-613/$31.00 ©2013 IEEE



(a) Compositional hierarchy

(b) Sharing structure

Fig. 3: Compositional hierarchy (a) and corresponding sharing structure (b) of our part-based Road Model.

variables.

Following [12], [13], [11], we use a mixture of L Gaussian kernels to approximate our potential functions. Accordingly, potentials are defined by

$$\psi_{ij}(\mathbf{x}_i,\mathbf{x}_j) \propto \varepsilon^0 \mathcal{N}_0(\mathbf{x}_j;\mu_0,\Lambda_0)+ \\ (1-\varepsilon^0)\sum_{l=1}^{L} w_{ij}^l \mathcal{N}(\mathbf{x}_j;\Omega_{ij}^l(\mathbf{x}_i),\Lambda_{ij}^l) \tag{2}$$

where ε^0 is a fixed outlier probability, which allows us to handle occlusions. The outlier distribution $\mathcal{N}_0$ is parametrized to be approximately uniform.

Each mixture component of our potential function has an assigned weight w_{ij}^l, mean μ_{ij}^l and covariance matrix Λ_{ij}^l. The covariance matrix Λ_{ij}^l represents uncertainties regarding the spatial relationship, between model parts $\mathbf{x}_i$ and $\mathbf{x}_j$. $\Omega_{ij}^l(\mathbf{x}_i)$ is a transformation function describing the spatial relationship of $\mathbf{x}_i$ and $\mathbf{x}_j$. In case of the potential $\psi_{1,17}(\mathbf{x}_1,\mathbf{x}_{17})$ depicted in Fig. 3a $\Omega_{ij}^l(\mathbf{x}_i)$ is defined by

$$\Omega_{1,17}^l = \begin{pmatrix} x_1 \\ y_1 \\ \alpha_1 \\ w_1 \\ l_1 \end{pmatrix} + \begin{pmatrix} \cos(\alpha_1) & 0 & 0 & 0 & 0 \\ 0 & \sin(\alpha_1) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} d_{17}^l \tag{3}$$

where d_{17}^l refers to an a priori patch-width, which e.g. corresponds to a common lane-width.

One challenge we were faced with formulating the potential functions is, that at least on dimension of the random variables represents a periodic variable $\alpha_i \in [0,2\pi)$. In order to overcome issue regarding the choice of origin, commonly von Mises distributions are employed.

Despite the fact that, non-parametric belief propagation can be applied to graphs containing von Mises distributions, the necessary modifications lead to additional model complexity. Hence, we use a linearized approximation introduced in [14] and [15] to model densities of periodic variables.

3) INFERENCE AND PART SHARING: Essential for the application of our approach is the ability to perform inference in real-time. Towards this goal we employ belief propagation (BP), which allows us to efficiently perform inference in our graphical model. In BP, messages $m_{ij}(\mathbf{x}_j)$ are passed from node i to j, encoding which state node j is in. Messages $m_{ij}(\mathbf{x}_j)$ can be computed iteratively using

$$m_{ij}(\mathbf{x}_j) = \int \psi_{ij}(\mathbf{x}_i,\mathbf{x}_j)\phi_i(\mathbf{x}_i,\mathbf{y}_i) \\ \times \prod_{k\in\mathcal{P}(i)\setminus j} m_{ki}(\mathbf{x}_i)d\mathbf{x}_i \tag{4}$$

where $\mathcal{P}(i)$ is the set of neighboring nodes of $\mathbf{x}_i$. Generally, we are interested in estimating the distribution $p(\mathbf{x}_j|\mathbf{y}_j)$ often referred to as the belief $b(\mathbf{x}_j)$ of a node $\mathbf{x}_j$. This belief can be determined by combining all incoming messages with the local observation potential as follows

$$b(\mathbf{x}_j) = p(\mathbf{x}_j|\mathbf{y}_j) \propto \phi_j(\mathbf{x}_j,\mathbf{y}_j)\prod_{k\in\mathcal{P}(i)} m_{ij}(\mathbf{x}_j) \tag{5}$$

If the potentials $\psi_{ij}(\mathbf{x}_i,\mathbf{x}_j)$ and $\phi_i(\mathbf{x}_i,\mathbf{y}_i)$ are Gaussian the belief can be calculated exactly using Eq. 4 and 5 [16]. However, in our case the potentials are multimodal due to noisy and ambiguous sensory cues. In this case, standard BP is not applicable.

In order to overcome this issues, we employ non-parametric belief propagation (NBP) [17], [12], which is a generalization of the particle filter [18] to general graphical models. In NBP messages are approximated by a set of L smoothed samples as follows

$$m_{ij}(\mathbf{x}_j) = \sum_{l=1}^{L} w_j^{(l)} \mathcal{N}(\mathbf{x}_j;\mu_j^{(l)},\Lambda_j) \tag{6}$$

where each mixture component l has an associated weight $w_j^{(l)}$, a mean $\mu_j^{(l)}$ and a bandwidth parameter Λ_j.

Using approximated messages according to Eq. 6 during BP leads to high computational complexity, because calculating the belief $b(\mathbf{x}_j)$ now requires the multiplication of several mixtures of Gaussians. In order to accelerate inference, we apply a sampling based approximation of the message products, which are similar to the one introduced in [19].

Another aspect that we can use to accelerate inference is that

978-1-4799-2914-613/$31.00 ©2013 IEEE

evidence is exclusively injected into our model from the observable nodes y_j. This allows us to divide message passing into an upwards sweep passing messages from the leaf nodes to the root nodes and a downwards sweep passing messages back from the root to the leaves. While message passing all nodes except the root and leave nodes receive messages from their children containing local evidence and messages from their parents ensuring the overall structure. Since our model depicted in Fig. 3a, contains many equal nodes messages sent from children (e.g. $\mathbf{x}_1^l$, $\mathbf{x}_3^l$ and $\mathbf{x}_5^l$) to their parents (e.g. $\mathbf{x}_{17}^P$, $\mathbf{x}_{18}^P$ and $\mathbf{x}_{19}^P$) contain equal information.

This circumstance can be used to apply belief sharing [11]. The basic idea of belief sharing is to avoid redundant calculations by combining random variables, which contain equal information during the upwards sweep. As can be seen in Fig. 3b we can combine the patch-nodes $[\mathbf{x}_{17}^P, \mathbf{x}_{18}^P, \dots, \mathbf{x}_{20}^P]$ and $[\mathbf{x}_{21}^P, \mathbf{x}_{22}^P, \dots, \mathbf{x}_{24}^P]$ into the nodes $\mathbf{x}_{17/18/19/20}^P$ and $\mathbf{x}_{21/22/23/24}^P$. The belief of e.g. node $\mathbf{x}_{17/18/19/20}^P$ is then shared between the nodes $\mathbf{x}_{25}^L$, $\mathbf{x}_{27}^L$ and $\mathbf{x}_{29}^L$. Consequently, we have to calculate the belief of node $\mathbf{x}_{17/18/19/20}^P$ only once, while we had to calculate the belief of nodes $\mathbf{x}_{17}^P$,$\mathbf{x}_{18}^P$,$\mathbf{x}_{18}^P$ and $\mathbf{x}_{20}^P$ individually using the compositional hierarchy.

4) SCENE REPRESENTATION: Besides reasoning about the location and configuration of objects in the scene the ability to reason about topological scene properties (e.g. number of lanes or road-classes) is of high importance for a complete scene understanding. Thus, we extended our part-based model as depicted in Fig. 4. Note that, to reduce the complexity of our illustration, we disregard the length of lanes.

As can be seen, the sensory evidence is provided by a road edge detector and a lane-marking detector on level 1. This evidence is then shared between different patch-types on level 2. Patches on level 2 differ in their width and their corresponding sensory cues. Different sensory cues are needed, because e.g. lanes on highways are usually bounded by two lane-markings, while urban roads are often bounded by a lane-marking on one side and a curbstone on the other side. Additionally, patches differ in their a-priori lane-width (see Eq. 3), which allows us to reason about e.g. if we are in a highway or an urban scenario.

The different patches are then shared between more complex objects on level 3 to level 6. Applying this part-based scene model allows us to reason about a variety of complex urban and non-urban scenarios, which greatly extends the area of application. Additionally, the high-level nodes introduce important contextual information and ensure the overall compatibility of parts.

One drawback of performing high-level reasoning in such complex part-based models while relying on error-prone sensory cues is that performing inference in real-time becomes a challenge. In order to make inference traceable, we propose a novel sequential message passing scheme.

5) DEPTH-FIRST MESSAGE PASSING: During the upwards sweep of BP, messages are passed from child nodes to all of their parents. Thus, each level of our model is processed one by one (see Fig. 4). Intuitively, this procedure can be understood as a breadth-first search for valid hypotheses in our graphical model. The main drawback of this procedure is that low-level nodes often contain many invalid hypotheses due to noise and clutter. Accordingly, while belief propagation messages passed through the graph contain many invalid hypotheses, which cause unnecessary computational complexity.

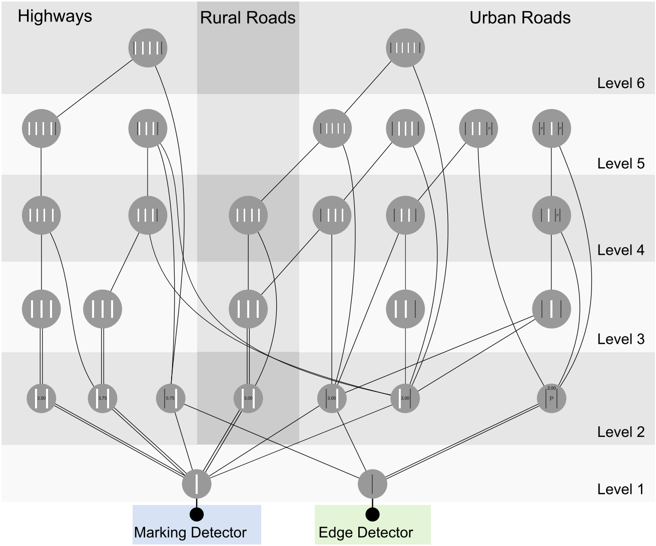


Fig. 4: Sharing structure applied to scene understanding. Levels represent objects of different complexity.

In order to improve the message passing, we propose a depth-first message passing scheme. The fundamental idea of depth-first message passing is to pass a single or a subset of samples in several sweeps from low to high-level nodes. Samples for message passing are randomly selected according to their weight, which gives us the ability to prefer valid low-level hypotheses. This is highly beneficial for the inference, because we first focus on low-level hypotheses, which are likely to be part of valid high-level hypotheses. This mean, we can accelerate the generation of valid high-level hypothesis while minimizing the propagation of invalid low-level. This in turn reduces the computational complexity of inference.

B. IMAGE EVIDENCE

Inspired by the virtual sensor concept proposed in [11], we use different 3d reconstruction approaches, which are applied to the same camera image in order to gain low-level information. The symmetrical local threshold method is used to detect lane-markings. According to the marking detector evaluation [20], it gives the best result in the general case. On the other hand, an edge detector is used to gain information in image regions where markings are missing, but road edges such as curbs are present. We employ a Sobel detector at multiple scales for a scale invariant edge detection.

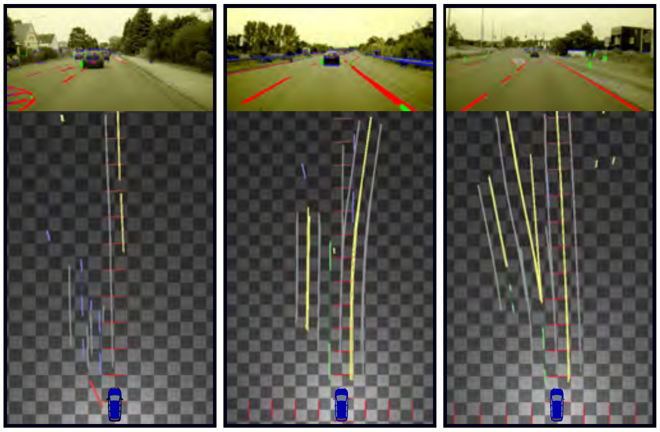


Fig. 5: Results of lane geometry estimation. (Top) Detected lane-marking features (red). (Bottom) Projection of the lane-marking detection into the vehicle reference frame (yellow, purple, green) and results of lane hypotheses generation (grey).

IV. RESULTS

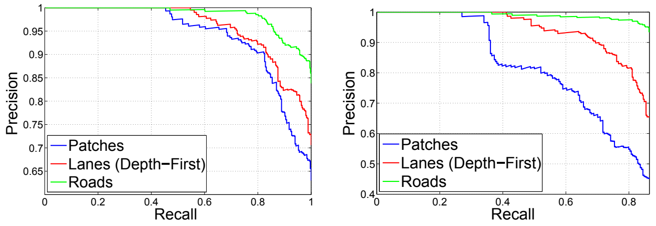
In this section, we present the results of applying our part-based model to several real world scenarios. Important aspects of our evaluation are (a) the accuracy with which lane and road geometries can be estimated, and (b) the recognition performance of our model. Another important aspect is (c) the computational performance, as it is an important aspect for in car applications. To show the performance of our approach in different environments, we tested our approach in highway, rural and urban scenarios (see Fig. 5).

The database used for this evaluation comprises several thousand individual video pictures of urban, rural and highway scenarios. In order to obtain ground-truth information, we use a high-accuracy navigation database, which contains an exact geometric lane description. During the evaluation, we align the database to the vehicles reference frame using a DGPS+IMU system. As data-input, we use the two visual cues presented in Sec. III-B. In respect to the available image resolution, we are able to detect features up to 80m (highway), 50m (rural) and 35m (urban).

The used part-based model has the same structure as depicted in Fig. 3. However, in order to cover the detection range of the used vision sensor, we extend the model by introducing random variables representing lanes composed of up to 40 patches. The root nodes of our graphical model represent roads with two or three lanes. For all tests we set the outlier probability ϵ^0 to 20% of the total likelihood. The a priori width and the length of the patches were set to w=3.5m and a length l=2.0m.

A. RECOGNITION PERFORMANCE

In order to illustrate the importance of contextual information incorporated by the high-level nodes, we compare the belief of patch locations estimated by our model with the ground-truth database. For this comparison, we approximate the true belief of patch locations by creating samples from



(a) Highway Scenario

(b) Rural Road Scenario

Fig. 6: Recognition performance after the upwards sweep (blue) and after incorporating lane (red) and road level (green) information.

our ground-truth database and estimating their weights using our observation potentials ϕ_j.

We estimated the marginal distribution over the patch locations in three different ways:

1) We only perform the upwards sweep of BP. Hence, the belief of patch-nodes only relies on sensory evidence.
2) In addition to the upwards sweep, we propagate messages from all lane nodes down to the patch-nodes. Thus, patch-nodes receive contextual information from their parents.
3) We perform a full bottom-up top down message passing scheme, which includes the road-nodes of our model.

As can be seen in Fig. 6 the recognition performance of our model increases drastically, as we incorporate contextual information.

This can be explained by the fact that patch-nodes perform inference over a relatively small area. Accordingly, they strongly rely on the presence of local, visual evidence. This means e.g. missing, occluded or damage lane-markings have a significant impact on the recognition performance.

Lanes, on the other hand, are based on a set of patches and thus combine sensory evidence from a larger area. Hence, the recognition performance is not as affected by missing local evidence as the one of patches.

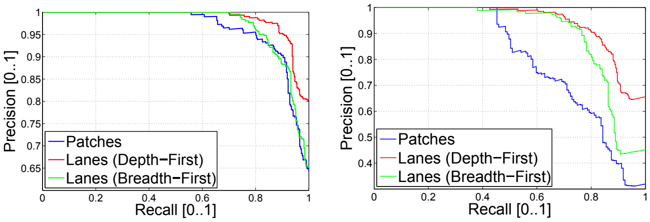
Incorporating contextual scene knowledge improves again the recognition performance of our model, as it ensures the overall compatibility of objects and object parts.

Some exemplary results for different scenarios are depicted in Fig. 5. Note hypotheses outside the lane-marking are supported by visual cues on one side, and by the outlier process on the other side. As a result, their associated weight is relatively low compared to hypotheses supported by two visual cues.

B. DEPTH-FIRST MESSAGE PASSING

As can be seen in Fig. 7, depth-first message passing outperforms standard breadth-first message passing in means of recognition performance, while using a significantly reduced sample set of only 25 samples. The reason for this major improvement is that by applying depth-first message passing we are mainly propagating high weighted samples. Hence, messages contain less invalid hypotheses than in breadth-first

978-1-4799-2914-613/$31.00 ©2013 IEEE



(a) Depth first and Breadth first Message Passing Highway

(b) Depth first and Breadth first Message Passing Road

Fig. 7: Recognition performance of depth-first (red) and breadth-first (green) message passing for different scenarios. In both scenarios depth-first message passing (25 samples) shows better performance than breadth-first message passing (150 samples).

message passing.

Furthermore, Tab. I shows that we can significantly reduce computational complexity by reducing the sample set, while achieving similar geometric accuracy. The overall performance of our approach is depicted in Tab. I, which shows that employing depth-first message passing allows us to apply our approach to real-time applications.

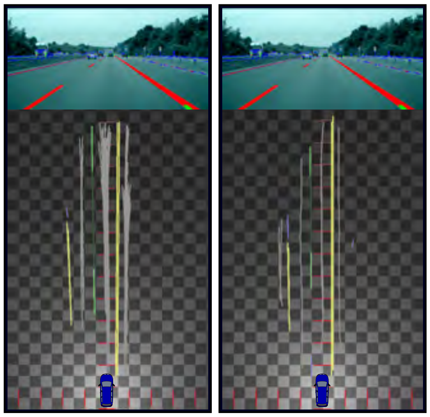


Fig. 8: Results of lane geometry estimation. (Top) Detected lane-marking features (red). (Bottom) results using breadth-first (right) and depth-first (left) message passing.

	Patches	Lanes (BF)	Lanes (DF)	Roads
time (ms)	1.47	75.29	4.41	5.38
RMS (m)	0.12	0.25	0.20	0.23

TABLE I: Computational time and location errors.

V. CONCLUSIONS

We presented a novel part-based approach to scene understanding in intelligent vehicles. Based on simple visual cues, our approach gives us the ability to robustly infer the geometry and layout of complex traffic scenes. Furthermore, we introduced a new depth-first message passing scheme that allows us to significantly reduce computational complexity, while performing inference. Our experimental results show that depth-first message passing significantly increases the performance of our approach, and allows us to perform inference in real-time.

In the future, we plan to add additional input sources to our model (e.g. vehicle tracks), allowing us to increase the performance in scenarios, where visual cues are not reliable. Furthermore, we believe that tracking hypothesis by adding temporal constrains to our model will greatly improve the robustness of our approach.

REFERENCES

[1] H. Wang, S. Gould, and D. Koller, “Discriminative learning with latent variables for cluttered indoor scene understanding,” *Computer Vision - ECCV*, pp. 435–449, 2010.

[2] A. Geiger, M. Lauer, and R. Urtasun, “A generative model for 3d urban scene understanding from movable platforms,” in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1945–1952.

[3] C. Wojek, S. Roth, K. Schindler, and B. Schiele, “Monocular 3d scene modeling and inference: Understanding multi-object traffic scenes,” *Computer Vision - ECCV*, pp. 467–481, 2010.

[4] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr, “Combining appearance and structure from motion features for road scene understanding,” 2009.

[5] D. Hoiem, A. Efros, and M. Hebert, “Closing the loop in scene interpretation,” in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[6] S. Bileschi, “Streetscenes: Towards scene understanding in still images,” DTIC Document, Tech. Rep., 2006.

[7] C. Wojek and B. Schiele, “A dynamic conditional random field model for joint labeling of object and scene classes,” *Computer Vision - ECCV*, pp. 733–747, 2008.

[8] A. Ess, T. Müller, H. Grabner, and L. Van Gool, “Segmentation-based urban traffic scene understanding,” in *Proceedings 20th British machine vision conference*, 2009.

[9] X. Wang, X. Ma, and W. Grimson, “Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[10] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari, “What’s going on? discovering spatio-temporal dependencies in dynamic scenes,” in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1951–1958.

[11] J. Spehr, D. Rosebrock, D. Mossau, R. Auer, S. Brosig, and F. Wahl, “Hierarchical scene understanding for intelligent vehicles,” in *Intelligent Vehicles Symposium*. IEEE, 2011, pp. 1142–1147.

[12] M. Isard, “Pampas: Real-valued graphical models for computer vision,” in *Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2003, pp. I–613.

[13] M. Isard and A. Blake, “Condensation-conditional density propagation for visual tracking,” *International journal of computer vision*, vol. 29, no. 1, pp. 5–28, 1998.

[14] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky, “Distributed occlusion reasoning for tracking with nonparametric belief propagation,” *Advances in Neural Information Processing Systems*, vol. 17, pp. 1369–1376, 2004.

[15] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, “Tracking loose-limbed people,” in *Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, june-2 july 2004, pp. I–421 – I–428 Vol.1.

[16] Y. Weiss and W. Freeman, “Correctness of belief propagation in gaussian graphical models of arbitrary topology,” *Neural computation*, vol. 13, no. 10, pp. 2173–2200, 2001.

[17] E. Sudderth, A. Ihler, M. Isard, W. Freeman, and A. Willsky, “Nonparametric belief propagation,” *Communications of the ACM*, vol. 53, no. 10, pp. 95–103, 2010.

[18] A. Doucet, N. De Freitas, N. Gordon *et al.*, *Sequential Monte Carlo methods in practice*. Springer New York, 2001, vol. 1.

[19] A. Ihler, E. Sudderth, W. Freeman, and A. Willsky, “Efficient multiscale sampling from products of gaussian mixtures,” *Advances in Neural Information Processing Systems*, vol. 16, pp. 1–8, 2004.

[20] T. Veit, J. Tarel, P. Nicolle, and P. Charbonnier, “Evaluation of road marking feature extraction,” in *11th International Conference on Intelligent Transportation Systems*. IEEE, 2008, pp. 174–181.