Unified situation modeling and understanding using hierarchical graphical Model
Einheitliche Situationsmodellierung und –Interpretation basierend auf grafischen Modellen

P. Pekezou Fouopi*, **G. Thomaidis**, **S. Knake-Langhorst**, **F. Köster**
Deutsches Zentrum für Luft- und Raumfahrt e.V.; Institut für Verkehrssystemtechnik, Braunschweig
{paulin.pekezoufouopi, georges.thomaidis, sascha.knake-langhorst, frank.koester}@dlr.de

Kurzfassung

Die Komplexität der Szene macht die Situationsmodellierung und –Interpretation schwierig. Um diese Aufgabe zu lösen, benutzen wir in dieser Arbeit das erweiterte Sensordatenfusionsmodell von *Joint Directors of Laboratories*. Ein probabilistisches graphisches Modell zur Modellierung der Kontextabhängigkeiten (örtlich, zeitlich, semantisch) zwischen den Objekten (Fußgänger, Fahrzeuge, Fahrspuren) und der Szene wird vorgeschlagen. Ein Inferenzalgorithmus zur Schätzung der Wahrscheinlichkeitsdichtefunktion an jeden Knoten des Graphs wird entworfen und implementiert.

Abstract

Situation modeling and understanding is a challenging task due to the scene complexity. In this work we use the extended *Joint Directors of Laboratories* model for sensor data fusion to solve this challenging task. We propose a non-directed graphical model to represent the global and local contextual dependencies (spatial, temporal, and semantic) between the objects (pedestrians, vehicles, and lanes) and the scene. We develop an inference algorithm to estimate the probability density function at each node of the graph in a bottom up top down approach using non-parametric belief propagation (NBP) scheme. The inferred objects are contextually consistent with respect to others objects and the scene.

1 Introduction

Advance driver assistance systems (ADAS) get increasing importance because they improve safety and comfort of traffic participants. Furthermore, they can be used to make traffic more efficient and ecological [1]. Some useful ADAS tasks like vehicle/pedestrian detection or ego vehicle free space estimation are developed separately or are combined without taking into account a priori relationships between entities (for example a vehicle normally does not move on a sidewalk) leading sometimes to false alarms. The *Joint Directors of Laboratories* (JDL) model introduced in [2] was initially developed for military applications. In [3] this model is considered as a conceptualization and common understanding of sensor fusion and revised for automotive applications. In this work we use this model and additionally introduce the process refinement as the part of the fusion process. In this process refinement part we propose a non-directed hierarchical graphical model to represent the global and local contextual dependencies (spatial, temporal, and semantic) between the objects (pedestrians, vehicles, and free space/lanes) and the scene. We develop an inference algorithm to estimate the probability density function at each node of the graph in a bottom up top down approach using a non-parametric belief propagation (NPB) scheme. The inferred objects are contextually consistent with respect to others objects and the scene.

The next sections of this work are divided as follows: In section 2, we present the related work, and describe our sensor data fusion model in section 3. Section 4 deals with our graphical modeling of contextual dependencies between the objects and the scene. In Section 5 we present some results, and conclude the work with section 6.

2 Related Work

The idea of using contextual information to jointly reason about the scene understanding tasks like object detection/categorization, semantic/geometrical scene labeling, saliency detection and scene categorization is motivated from the results of human visual perception research. In [4] five features that are important for human vision are proposed: support (objects should not be floating), interposition (objects should occupy different volumes), probability (objects should or should not appear in certain scenes or together), position (objects should appear in typical locations in the scene or relative to others objects), and size (objects have typical relative sizes) (see [5], [6], [7], [8], [9], [10], [11], and [12] for the implementation of this features). In our work we use different combination of these features.

A critical survey of context based object categorization is presented in [13]. The author shows how global and local context information can be combined with pixels, regions, and objects' appearance to improve the object categorization. The "gist" descriptor proposed in

[14] is one of the most popular global context features for predicting the scene class of a given image or the excepted objects as well as their positions and scales in an image (see [5], [6], [12], and [9]). We plan to use this descriptor in our work.

To combine the contextual information graphical models are more suitable than classifier, because they can encode the local dependencies in the scene. These dependencies are used to make a globally consistent prediction. The main disadvantage of combining this information is the complex and expensive computations [13]. One of the most used graphical models is the Conditional Random Field (CRF). This model is used for e.g. to jointly reason about the object detection, the scene classification (see [12]) and the semantic scene segmentation (see [15]). A similar model is also used to estimate the scene geometry (see [16]) or for the object categorization (see [7]). The main differences between authors using graphical models are the methods for learning (e.g. Expectation Maximization [17], Maximum Likelihood [5] or Maximum a posteriori Probability [10] estimator) and inference (e.g. Max-Product Belief Propagation [18], particle-based NBP [8] or Metropolis-Hastings [19] algorithm). Our approach differs from those, since we use a non-directed hierarchical graphical model to represent the scene. Object nodes are pairwise combined modeling scene parts or the whole scene. We develop an inference algorithm to estimate the probability density function (PDF) at each node of the graph in a bottom up top down approach using NBP.

Instead of using graphical model the author in [20] propose a two layer classifier model, where the outputs of the first layer classifiers are used to improve the second layer and vice versa. The system can jointly improve tasks like depth map, scene geometry and saliency estimation as well as object detection. A hybrid approach is proposed in [21] where the author models the graphical model nodes as classifier and learns the message passing inference. The resulting system is a sequence of predictors, which is used to classify 3D point cloud and estimate 3D surface.

There are many levels where different tasks can be combined together. Authors in [8], [17], und [18] combine low level pixels and/or regions features and introduce contextual information to jointly resolve the different tasks. Others authors take as input the independent detector outputs (object, scene class, scene geometry, etc.) and just model high level dependencies to improve each output (see [7], [20], and [22]) or to resolve new task (see [12], [15], and [5]). We follow the second approach in our work. A combination of low and high level information is also possible (see [19], and [23]).

3 System Overview

Figure 1 shows the overview our sensor data fusion system. As mentioned above we use the extended JDL model as described in [24] (level 0 to 4). All tasks on the levels 0 and 1 are either independently processed or weak coupled in a bottom up approach. For each output we give assumption about the content and the quality (see the two black arrows). The level 1 outputs a list of detected objects like pedestrians, vehicles, and Occupancy Grid Map (OGM). We model the contextual dependencies between the objects and the scene in the "*Situation Modeling and Interpretation*" (SMI) part using a graphical model. This joint reasoning allows us to improve the "*Object Detection & Tracking*" and even the "*Sensor Data Processing*" part with respect to the scene consistency (see violet arrows) in a top down approach. The next paragraphs will deal with the detailed description of this system. For level 0 and 1 we just implement state of the art methods, because this levels are just the inputs our SMI task.

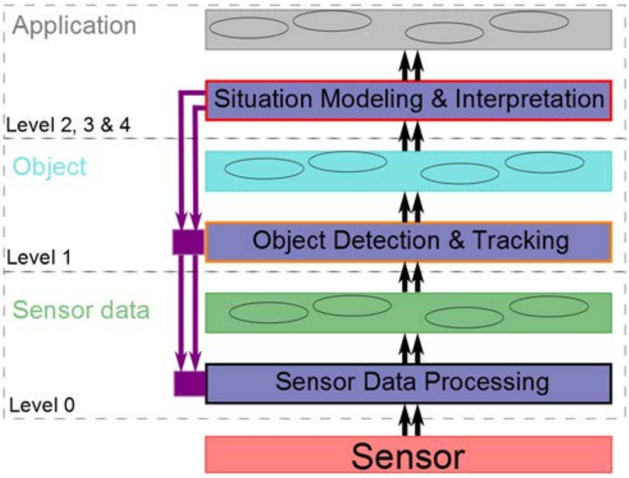


Figure 1: Layer architecture our sensor data fusion system. The hierarchical graph is modeled in the "*Situation Modeling & Interpretation*" part. The violet arrows represent the feedback loop and the black one the contents and quality. For more details about level 0 to 4 see [24].

3.1 Level 0: Sensor Data Processing

The laser scanner system used in our experimental vehicle consists of four Sick LUX laser scanner sensors which are arranged on the test vehicle's front and rear. Each sensor has a detection field 90^o degrees and 120 meters in azimuth and range respectively. The transmitted laser pulses are reflected by objects within the measuring range. These echo pulses are delivered by the sensor as a scan point.

The stereo camera system consists of two gray digital cameras mounted ahead on the ego vehicle roof. We use a stereo base width of 46 cm and a focal length of 1281pixel allowing us to have good depth resolution up to 40 m in front of the ego vehicle. The Stereo Data

Processing outputs a dense depth map using a GPU implementation of the Semi Global Matching (SGM). The SGM is described in [25].

3.2 Level 1: Object Detection and Tracking

In this section we describe the OGM computation, the pedestrian and vehicle detection which are the most important objects in our work.

3.2.1 OGM computation

The occupancy grid (OG) representation [26] and the related SLAM and DATMO algorithms are a widely researched topic in bibliography [27], so in the current work the mere basics of the framework is presented.

In the OG representation, the vehicle environment is divided into a two-dimensional lattice m of rectangular cells (extensions however exist for polar coordinate systems). Each cell is assigned with a value indicating the probability that the cell is occupied by an obstacle. A high value of occupancy grid indicates the cell is occupied and a low value means the cell is free.

In general the mapping problem from a moving robot is summarized as follows:

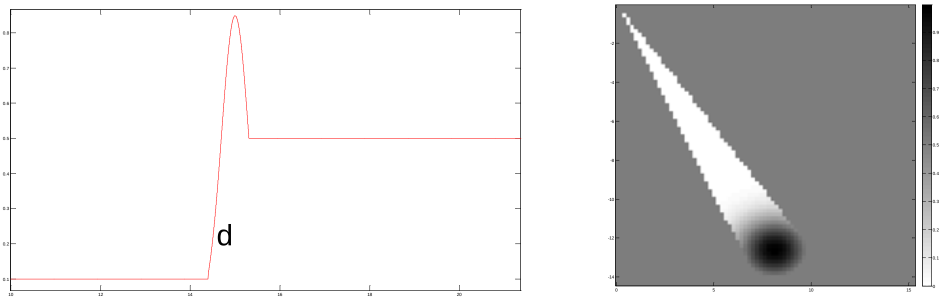
$$p(s_t, m, | z^t, u^t) = \eta \, p(z_t | s_t, m) \int p(s_t | u_t, s_{t-1}) \, p(s_t, m, | z^{t-1}, u^{t-1}) ds_{t-1} \qquad \textbf{3.1}$$

, where s is the robot pose, m is the map, z_t is the sensor measurement taken at time t, and u_t specifies the robot motion command asserted in the time interval $[t-1, t)$. Values notated by a superscript ^{t} refer to all data leading up to time t. As we see from the above equation, a slam algorithm estimates both the map and the robot pose.

In general a Bayes filter is used to calculate the posterior over the occupancy of each grid cell. A grid cell with coordinates $\langle x, y \rangle$ and occupancy m_{xy}, has the posterior probability calculate as follows using the log odds form [28]:

$$\log \frac{p(m_{x,y} | z^t, s^t)}{1 - p(m_{x,y} | z^t, s^t)} = \log \frac{p(m_{x,y} | z_t, s_t)}{1 - p(m_{x,y} | z_t, s_t)} + \log \frac{p(m_{x,y} | z^{t-1}, s^{t-1})}{1 - p(m_{x,y} | z^{t-1}, s^{t-1})} \qquad \textbf{3.2.}$$

As we can see from the above the equation, only the calculation of the probability density $p(m_{x,y} | z_t, s_t)$ is required, which is called the inverse sensor model. The inverse sensor model can be retrieved from the working principle of the sensor. Assuming we have a laser reflection at distance d (see Figure 2a), we assume that space until the reflection is free and the space behind the reflection is unknown since it is occluded by the source of the reflection. A 2D and 3D representation of the inverse sensor model are shown in Figure 2a, and 2b.



(a) (b)

Figure 2: The 2D (a) and 3D (b) representation of the inverse sensor model.

Pose estimation or localization is another important aspect in the mapping problem, since accurate robot (in this work ego vehicle) position is required in order to accurately update the map. Many methods are proposed and there is an extensive bibliography in this topic.

In our work we first compute two separate OGMs based on the LUX and the stereo camera system and then fuse it into one OGM.

3.2.1.1 OGM with LUX

The algorithm architecture is presented in Figure 3. The first step of the algorithm consists of determining the vehicle movement between two consecutive scans. The alignment between the two scans is achieved with the use of the iterated closed point (ICP) algorithm [29]. The output of the algorithm is the estimation of the vehicle displacement and rotation which is used to transform the previously estimated grid into the new vehicle pose. In this iterative way we can estimate the new robot pose s_t based on the previous pose s_{t-1} and the robot estimated motion and rotation, $[\Delta x \quad \Delta y]_{[t-1t)}$ $\theta_{[t-1t)}$ respectively [30]. Then each cell's posterior probability $p(m_{x,y} \mid z^t, s^t)$ given the measurements z^t and poses s^t up to time t are updated by applying the inverse sensor model $p(m_{x,y} \mid z_t, s_t)$ and using the *log odds* representation of Bayes' rule as explained in the previous section. The estimated OGM is shown in Figure 5b.

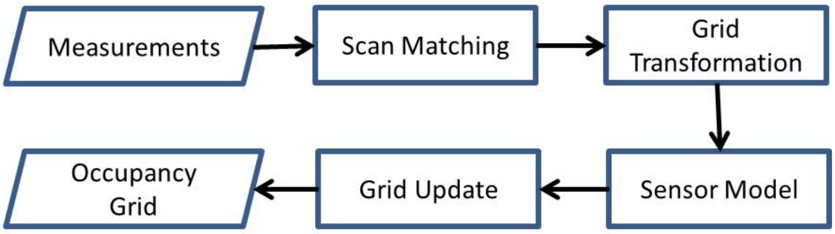


Figure 3: Block diagram of the SLAM algorithm used for the Laser scanner grid update.

3.2.1.2 OGM with stereo camera system

Detection of free space/obstacles using stereo system can be made in the 3D Euclidian or the v-disparity space. These two approaches are compared in [31].

The v-disparity space approach, originally proposed in [32], computes the v-disparity space accumulating the pixels with the same disparity value from a given image line. Finding the ground plane in the image is equivalent to find a line in the v-disparity space under the assumptions that the ground plane is flat, one of the most seen parts in the image, and that the camera roll angle is small enough to be neglected. Usually Hough transformation is used to detect the ground plane line. This line equation depends only on the parameters camera pitch angle and height above the ground plane. A threshold is used to separate free space from obstacles pixels according to their distance to the ground plane line. While this approach is easy to compute, and achieve good estimation, when the road is flat (e.g. highways environments), it suffers from wrong detection when the assumptions mentioned above are violated (e.g. non-flat road in urban scene or ground plane as non-dominant part in the image due to obstacles) [31]. [33], [34], and [35] extend this approach to overcome the problems mentioned above. We will not go more in details because this is out of scope our work. We just implement the original approach proposed in [32].

The 3D Euclidian approach computes the 3D map of the image and either uses a fitting algorithm like least square in combination with voting schema like RANSAC or propagation algorithm to estimate the ground plane. In both cases a set of start points corresponding to the road and the road equation have to be chosen. Smoothness assumption and prior knowledge about road slope are integrated to improve the result. The main advantage is that non-flat road can be better estimated. This approach is more sensitive to 3D stereo computation errors and is time consuming. Also the choice of the start road points could be difficult if for example the road is occluded. Following this approach the author in [36] models the road as a quadratic surface. Using a density and a Digital Elevation Map (DEM) road start points are selected and the road plane is fitted using RANSAC. The set of road points are extended with new points based on smoothness constraint and the road plane is fitted again. Stereo computation errors are taken into account. This method can separate the road from traffic isle and the other obstacles. We implement a simplified version of this approach in our work and just propagate the points in the DEM with respect to the smoothness between neighbor points. The result is shown in Figure 5c.

3.2.1.3 OGM Fusion

In our approach, a separate occupancy grid is built for each sensor and the two grids are fused on a cell level. Since the two grids have the same dimensions and cell size, the grids fusion is a matter of combining the occupancy probabilities of each cell. Assuming the independence of the laser scanner and stereo system measurements, the fused occupancy estimate of each cell can be computed using the Bayes Theorem ("Naive" Bayesian Fusion) [37].

$$p(m_{x,y} \mid z^t)_{FUSED} = \frac{p(m_{x,y} \mid z^t)_{Laser}\; p(m_{x,y} \mid z^t)_{StereoCam}}{p(m_{x,y} \mid z^t)_{Laser}\; p(m_{x,y} \mid z^t)_{StereoCam} + \left(1 - p(m_{x,y} \mid z^t)_{Laser}\right)\left(1 - p(m_{x,y} \mid z^t)_{StereoCam}\right)}$$

3.3.

3.2.2 Vehicle Detection

Vehicle detection is an important task for driver application. One application is the adaptive cruise control (ACC) where the system is able to automatically adapt the ego vehicle distance and velocity avoiding collision with the front vehicle. Vehicle detection using vision systems remains a challenging task due to the huge appearance variability [38]. In [38] the author reviews different vehicle detection approaches and describes the main tasks of hypotheses generation and validation.

The hypotheses generation consists of generating regions of interest where possible vehicles are located in the image using a sliding windows approach. On this step prior knowledge like symmetry [39], color, edge, corner, texture, depth [40], aspect ratio, Stixel [39] and motion are used. Also contextual information like contact point to the ground plane (vehicles usually move on the road/ground plane) can be integrated.

The hypotheses validation step classifies the generated hypotheses as vehicle or not using template-based or appearance-based approach. The template-based approach estimates the correlation between the generated hypotheses and predefined patterns (e.g. "U" shape, license plates, and "moving edge closure", etc.). Vehicle hypotheses are validated if the correlation is bigger than a threshold. Appearance-based approach use features like SIFT, Gabor Filter, Histogram of Oriented Gradients (HOG) [41] or Haar Wavelet [40], and classifier like Support Vector Machines (SVMs) [41], neural networks, AdaBoost [40] or Bayes classifier to classify the generated vehicle hypotheses. Due to the features variation with vehicle viewing point, classifiers are trained for different viewing point like front, back, front-side, and back-side (see [40], [41] and [42]). Advantages and disadvantages of these approaches are discussed in details in [38]. In this work we generate vehicle hypotheses using the detected ground plane and the “u-v”- disparity space. We use SURF [43] features and “bag of words/key points” [44] to train a SVM classifier.

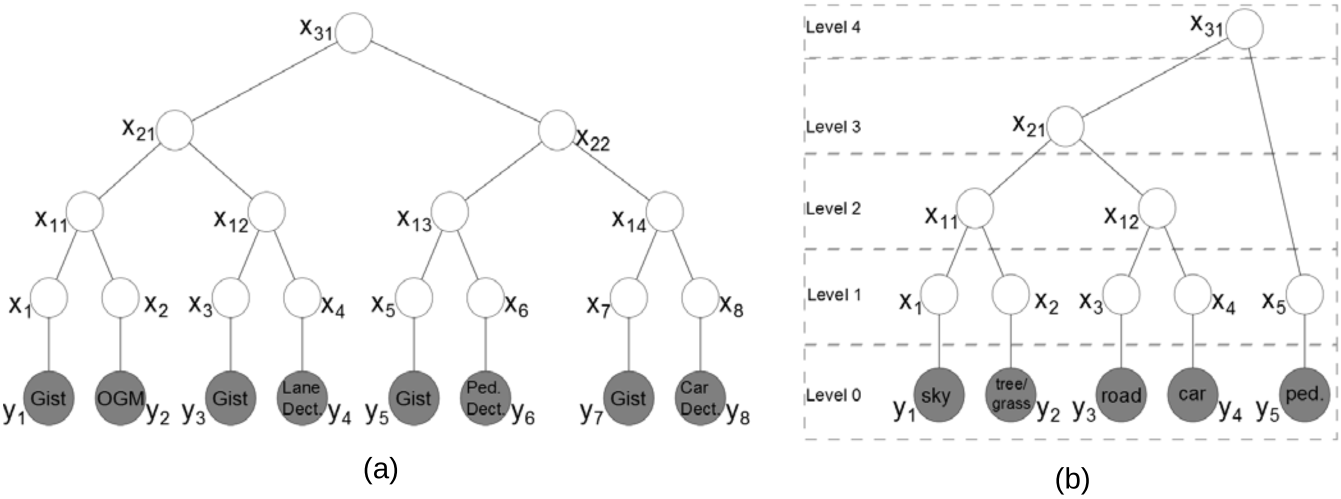


Figure 4: Our tree based hierarchical graphical model for scene modeling and understanding. Shaded nodes are observed in the learning and inference steps. White nodes are objects, scene parts or the whole scene. These hidden nodes are just observed in the learning step and inferred resulting in a contextual consistent scene. The left tree (a) is a generic one while the right (b) is learned using data from [7].

3.2.3 Pedestrian Detection

Pedestrian detection is, like vehicle detection, a challenging task. Many approaches are proposed in the literature. We refer to the comparison proposed in [45] for more details.

Similar to vehicle detection hypotheses are generated using prior knowledge like aspect ratio, Stixel [46] and contextual information (pedestrian usually stay on the ground plane) in a sliding window approach.

The hypotheses validation is either shape or features based or combined [47].

Shape based approach [47] use prior knowledge about pedestrian to generate shapes set. A given shape is matched with this set using a correlation function and classified as pedestrian or not with respect to a threshold.

Features like HOG [48], Local Binary Pattern (LBP) [49], color, texture or motion [50] are combined with classifier like SVM [48] or AdaBoost [51] to predict the class (pedestrian or not) of a given image window. In this work we use the OpenCV implementation for people detection similar to [48] and we add the hypotheses generation step using Stixel and aspect ratio.

4 Level 2, 3 & 4: Situation Modeling & Interpretation – An unified approach

As mentioned above the idea of holistic scene understanding is motivated by human visual perception. Low and high level features are combined in a bottom up top down approach to jointly reason about scene information like saliency, geometry, and objects with respect to contextual dependencies. In the next step we will propose a graphical model for contextual dependencies. Learning and inferring the model parameters will be explain.

4.1 Graphical model for contextual dependencies

Inspired from the graph theory, probabilistic graphical models $G = (V, E)$ are defined as a set of vertices (nodes) $i \in V$ associated with random variables $x = \{x_i\}_{i=1}^{n}$ and edges $(i, j) \in E | i, j \in V$ modeling the dependencies between $x_i \in x$ and $x_j \in x$. While estimating the joint probability over the set of random variables $x = \{x_i\}_{i=1}^{n}$ is complex graphical models allow us to factor the representation of this probability into modular components using the independence properties [52]. Undirected graphical models called Markov Random Field (MRF) are more suitable for our work allowing us to reason in a bottom up top down manner. Using pairwise MRF the joint probability [52], [53]

$$\boldsymbol{P}(\boldsymbol{x}) = \prod_{(\boldsymbol{i},\boldsymbol{j}) \in \boldsymbol{E}} \boldsymbol{\psi}_{\boldsymbol{ij}}(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{j}}) \prod_{\boldsymbol{i} \in \boldsymbol{E}} \boldsymbol{\psi}_{\boldsymbol{i}}(\boldsymbol{x}_{\boldsymbol{i}}) \qquad 4.1$$

is the product of the pairwise and single node potentials $\psi_{ij}(x_i, x_j)$ and $\psi_i(x_i)$ containing the dependencies between neighbor nodes and single-node constrains. Estimating the probability in equation 4.1 can be divided in 3 tasks: estimating the model structure, learning the potentials, and inferring the probability.

4.2 Tree-based hierarchical graphical model

In this work we use a tree-based model as showing in Figure 4a instate of a fully connected network reducing the model complexity. The leaf nodes are the object detection result combined with prior knowledge (in this case “gist”). These nodes are observed in the learning and inference steps. We focus on binary tree and learn the structure using the Algorithm 4.1.

Algorithm 4.1: Algorithm for learning the binary tree structure our hierarchical graphical model.

```
Input: Leafs nodes $V_0 = \{i\}_{i=1}^{n}$ corresponding to the n object classes
The co-occurrence matrix $M \in \mathbb{N}^{n \times n}$
Output: Output graph $G = (V, E)$
Initialization: $V \leftarrow V_0; E \leftarrow \emptyset; V_{prev} \leftarrow V_0;\ V_{crt} \leftarrow \emptyset$
Do
    If $|V_{prev}| = 1$ Then Return $G = (V, E)$ End if
    New node: $V_{crt} \leftarrow V_{crt} \cup \{i, j\} |\ i \neq j ;\ i, j \in V_{prev} ;\ \{i, j\} = argmax_{\ k,l} M(k, l)$
    New edge $E \leftarrow E \cup \{(i, j)\}$
    $V_{prev} \leftarrow V_{prev}\ / \{i, j\}$
While $|V_{prev}| \leq 1$
If $|V_{prev}| = 1$ Then $V_{crt} \leftarrow V_{crt} \cup V_{prv}$ End if
$V \leftarrow V \cup V_{crt};\ V_{prev} \leftarrow V_{crt};\ V_{crt} \leftarrow \emptyset$
If $|V_{prev}| > 1$ GO TO $G = (V, E)$ Else Return $G = (V, E)$ End if
End
```

This algorithm takes as input the detected objects and a co-occurrence matrix containing the pairwise objects occurrence. This matrix is learned using a set of labeled scene images. The resulting tree is a hierarchical representation of the scene. The root node contains all the scene objects and the children nodes scene parts. Edges between nodes with high co-occurrence are preferred. Using hierarchical model for scene understanding task is also proposed in [53], [17] and [54] but the authors use different approach to learn the tree.

4.3 Learning the potentials

In this work we have two pairwise potentials:

1. The co-occurrence potentials are learned using labeled scene as described above. We count the objects co-occurrence and save it into the co-occurrence matrix.
2. The spatial dependencies potentials are learned in a similar way. We use the discrete spatial relations *above*, *below*, *around*, and *inside* as proposed in [6] or [7]. For every object pair we count the occurred spatial relation using labeled scene images.

4.4 Inference in tree based probabilistic graphical model

The inference task consists of estimating the probability density function at each tree node. We use a particle-based NBP algorithm as proposed in [53] and [54]. The belief

$$\boldsymbol{b_i(x_i) = \sum_{l=1}^{L} w_i^{(l)}\, \mathcal{N}\left(x_i; x_i^{(l)}, \Lambda_i\right)} \qquad 4.2$$

is the weighted sum of Gaussian functions (mixture of Gaussians) and is approximated by extending equation 4.1 as follow:

$$\boldsymbol{b_i(x_i) = p(x_i|y) \propto \psi_i(x_i, y_i) \prod_{j\in\Gamma(i)} m_{ji}(x_i)} \qquad 4.3$$

combines the local observation potential $\psi_i(x_i, y_i)$ (where $y = \{y_i\}_{i=1}^{n}$ is the set of observable nodes) with the incoming message

$$\boldsymbol{m_{ji}(x_i) = \int \psi_{ji}(x_j, x_i)\psi_j(x_j, y_j) \prod\nolimits_{k\in\Gamma(j)/i} m_{kj}(x_j) dx_j} \qquad 4.4.$$

$\Gamma(i) \in V$ defines the neighborhood set of $i \in V$. We use a particle based representation of the message $m_{ji}(x_i)$ similar to equation 4.3. The inference algorithm starts on the observed leaf nodes and propagates the messages up to the root (bottom up). In the second step messages are propagated from the root to the leaf nodes resulting to scene consistent belief estimation on every node (top down).

5 Results

5.1 Evaluation of the OGM Fusion

In order to evaluate the performance of the OGM fusion, data recordings from of our experimental vehicle FASCar [55] driving in the city Braunschweig were used. In Figure 5, a representative frame is shown along with the OGM built by each sensor system independently and the fused OGM as well. As it can be seen in Figures 5b and 5c, the laser scanner OGM has less noise and better long range accuracy whereas the camera delivers more features (better vehicle shapes for example), in short distances. Also it seems useful not to account for camera detections in the extreme azimuth angles. By taking this into account the equation 3.3 was applied to fuse the two OGMs. The final fused OGM is shown in Figure 5d. As it can be seen, the long range advantage of the laser scanner is mainted whereas on the close range, the neighbouring vehicle contour is representated better in comparison with the laser OGM.

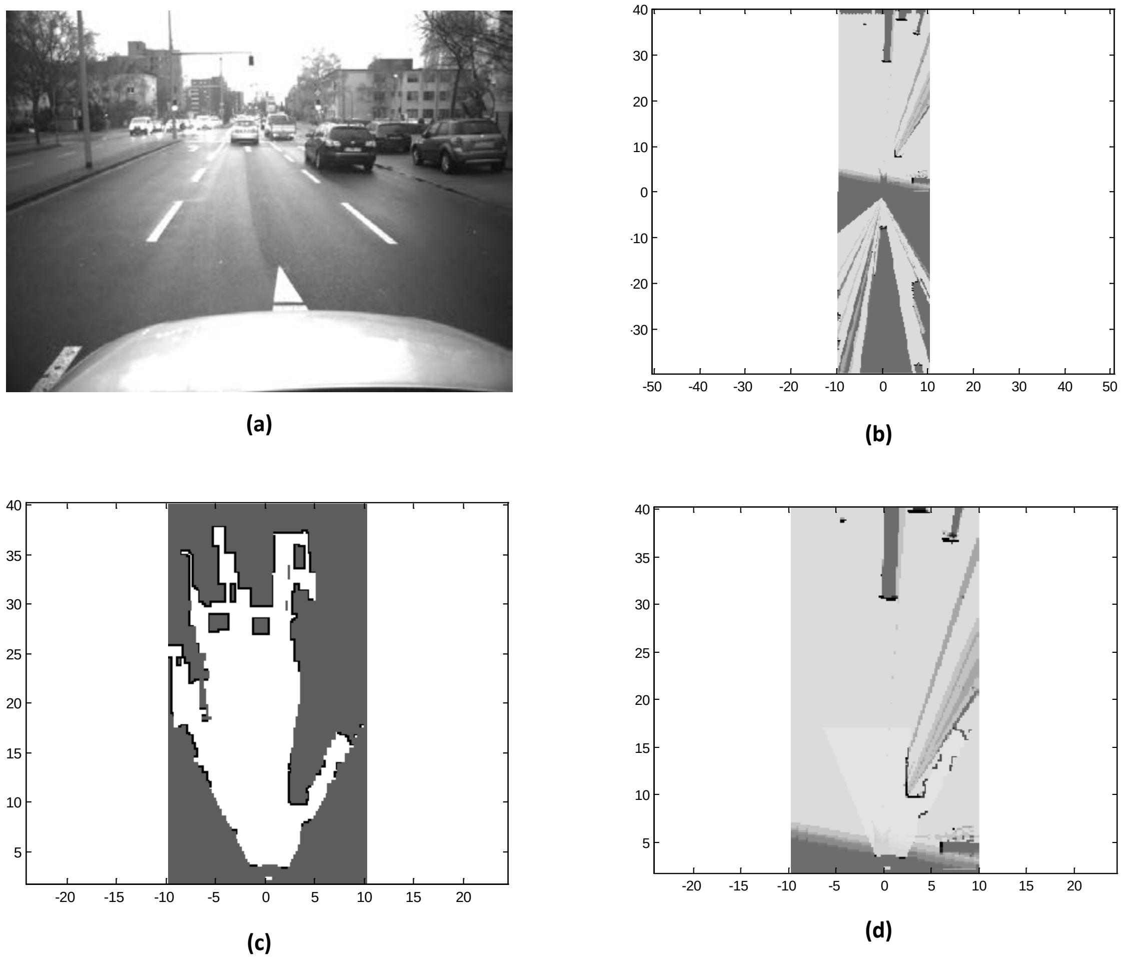


Figure 5: Explanation of the grid fusion scheme used. On (a) the camera frame used. On (b) the OGM built by the laser scanner measurements. On (c) the OGM build by the stereo system and finally on (d) the fused OGM. It can be seen that a direct fusion of cells would lead to degraded performance, whereas if the stereo OGM is taken into account for distances < 15 and angles that are not in the extreme ends, the end result adds more information about the scene.

5.2 Evaluation of the tree learning algorithm

To evaluate our tree learning approach we use the data proposed in [7] and focus on the objects *sky*, *tree/grass*, *road/*, *vehicle* and *pedestrian*. This is an extension of the objects showed in Figure 4a with more contextual information. We remove the "gist" nodes since this feature is yet not available. We use the co-occurrence matrix as input for Algorithm 4.1. Figure 4b shows the learned tree. The most co-occurred objects pairs *sky-tree* and *road-vehicle* are preferred in the hierarchical levels 2 and 3. *Pedestrian* are just integrated in the last level (level 4). This tree corresponds to our idea of building nodes with strong contextual information in low hierarchical levels resulting in a rapidly consistent scene representation during the bottom up inference step.

5.3 Evaluation of the inference on the tree-based model

We apply our inference algorithm on the tree-based model in Figure 4b in a bottom up top down approach as proposed in [53].

In the bottom up part we use the object detector outputs (level 0) as inputs and run the following steps:

1. **Initialization**: On level 1 leaf nodes partial beliefs haven't incoming messages and just depend on the observed nodes potentials $\boldsymbol{\psi_i}(\boldsymbol{x_i}, \boldsymbol{y_i})$ (detector outputs). These beliefs are computed using equation 4.2 where $w_i^{(l)}$ is the detection score for the object (region) l.

$$\boldsymbol{x}_i^{(l)} = \left(\boldsymbol{c}_u^{(l)}, \boldsymbol{c}_v^{(l)}, \boldsymbol{h}^{(l)}, \boldsymbol{b}^{(l)}\right) \in \mathbb{R}^4 \qquad 5.1$$

   is the region bounding box. The covariance matrix $\Lambda_i$is set using experimental result.

2. **Message sending**: Sending a message from level 1 to 2 consists on generating samples as described in the following example: suppose we want to send a message from node x_1 and to x_{11} where the partial belief $b_{1/11}(x_1)$ is known and represented as a mixture of Gaussians. For each object bounding box $x_i^{(l)}$ (see equation 5.1) we generate $k \in \mathbb{N}$ new samples from the uniform distribution $U(k; a, b)$. The interval $[a, b]$ corresponds to the lower and upper bound of a region in the image with respect to the learned spatial dependencies between nodes x_1 und x_2 object classes since we are just interested on the vertical location (horizontal information can be neglected for the used data [7]). It means each *sky* object votes for a set of probable *tree/grass* objects vertical position, according to the learned spatial dependencies and vice versa. We weight the new samples multiplying $w_i^{(l)}$ with the learned co-occurrence frequency. We apply this example for other nodes between level 1 and 2.

3. **Partial beliefs update:** in this section we also use the example of node x_{11} to explain this step. Since the incoming messages $m_{111}(x_{11})$ and $m_{211}(x_{11})$ were computed in the previous step, the partial belief $b_{11/21}(x_{11})$ is just the product of the incoming messages since this node is not observed (see equation 4.3). The result is a mixture of Gaussians where the mean vector

$$\boldsymbol{x}_{11}^{(l)} = \left(\boldsymbol{x}_1^{(l)}, \boldsymbol{x}_2^{(l)}\right) \in \mathbb{R}^8 \qquad 5.2$$

contains the updated pair *sky* and *tree/grass* bounding boxes. The weight $w_{11}^{(l)}$ is the product of the incoming message weights and a function of those messages contextual consistency. Consistent pairs will be preferred while inconsistency will be penalized. We update all the nodes in level 2 in a similar way and repeat the steps 2 and 3 until arriving the root node. The most weighted root belief sample is the most consistent scene configuration.

Starting on the root node we propagate the message on a top down approach to the leaf nodes following the steps 2 and 3 described above. The step 1 is skipped since the root node belief is known. The final belief on a children node (e.g. x_1)

$$\boldsymbol{b_1(x_1) \propto b_{1/11}(x_1) m_{111}(x_1)} \qquad 5.3$$

is approximated as a product of the partial belief computed in the bottom up approach above and the parent node incoming message. It means that on the leaf nodes detected objects are reweighted with respect to the scene consistency. Inconsistencies are then penalized.
Since we are now evaluating the output our approach concrete results and comparison with other approaches will be proposed in our next paper.

6 Conclusion

In this work we have addressed the problem of scene understanding and particularly object detection in a holistic manner. We presented our system architecture for sensor data fusion as an extension of the extended JDL model. Levels until "*Object Detection and Tracking*" are reviewed and state of the art detector are implemented. We have focused on the "*Scene Modeling & Understanding*" part. In this part we have proposed a tree based hierarchical graphical model to integrate contextual information with object detection resulting in a consistent scene representation. We have proposed an algorithm to learn the tree-based model and shown that this model simplifies the fully connected model without losing strong scene context information. We have used the object co-occurrence and spatial relation to learn the parameter of this model (potentials). We have proposed a particle-based NBP algorithm to infer the belief on each tree node in a top down bottom up approach. The inferred scene is contextually consistent and the object detection outputs are reweighed preferring consistent objects.
In our future work we will first finish the evaluation on the proposed method and compare it with the state of the art. Furthermore following points will be investigated:

1. Learning of tree-based model: Since this model simplifies a fully connected network, information is lost. We will look more in depth to see the effect on the output. We will analyze the algorithm complexity, compare it with another algorithm and improve it.
2. Extending the contextual information: We will integrate more contextual information (temporal, scale, and “gist”, etc.) and scene understanding output (saliency, geometry, parsing, segmentation, etc.).
3. Particle-based NBP: Since the complexity grows with the sample number during the message passing we plan to reduce this complexity using similar approach as proposed in [53] and [54]. We also want to investigate sampling method to approximate the messages and beliefs computation.
4. Context on low level: instate of reasoning on high level information some authors (see section 2) integrate context information on pixel or region level and jointly reason about many scene understanding task. Other Authors combine low and high level information. We will investigate these approaches.

[1] S. Nedevschi, R. Danescu, D. Frentiu, T. Marita, F. Oniga, C. Pocol, R. Schmidt and T. Graf, “High Accuracy Stereo Vision System for Far Distance Obstacle Detection,” *IEEE Conference on Intelligent Vehicles Symposium,* pp. 292-297, 14-17 June 2004.

[2] F. E. white, “A Model for Data Fusion,” *First National Symposium on Sensor Fusion,* 1988.

[3] A. Polychronopoulos and A. Amditis, “Revisiting JDL model for automotive safety applications: the PF2 functional model,” *9th International Conference on Information Fusion,* pp. 1-7, 10-13 July 2006.

[4] I. Biederman, R. J. Mezzanotte and J. C. Rabinowitz, “Scene perception: Detecting and judging objects undergoing relational violations,” *Cognitive Psychology,* vol. 14, no. 2, p. 143–177, April 1982.

[5] M. J. Choi, A. Torralba and A. Willsky, “A Tree-Based Context Model for Object Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 2, pp. 240-252, February 2012.

[6] M. J. Choi, A. Torralba and A. S. Willsky, “Context models and out-of-context objects,” *Pattern Recognition,* vol. 33, no. 7, pp. 853-962, May 2012.

[7] C. Galleguillos, A. Rabinovich and S. Belongie, “Object categorization using co-occurrence, location and appearance,” *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1-8, 23-28 June 2008.

[8] S. Kim and I. S. Kweon, “Visual Context-based Scene Interpretation in Indoor Environment,” *The 2nd International Conference on Ubiquitous Robots and Ambient Intelligence,* November 2005.

[9] C. Liu, J. Yuen and A. Torralba, “Nonparametric Scene Parsing via Label Transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 12, pp. 2368-2382, December 2011.

[10] M. Boutell, J. Luo and C. Brown, “Scene Parsing Using Region-Based Generative Models,” *IEEE Transactions on Multimedia,* vol. 9, no. 1, pp. 136-146, January 2007.

[11] S. Ross, D. Munoz, M. Hebert and J. Bagnell, “Learning message-passing inference machines for structured prediction,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2737-2744, 20-25 June 2011.

[12] A. Torralba, K. P. Murphy and W. T. Freeman, “Using the forest to see the trees: exploiting context for visual object detection and localization,” *Communications of the ACM,* vol. 53, no. 3, pp. 107-114, March 2010.

[13] C. Galleguillos and S. Belongie, “Context based object categorization: A critical survey,” *Computer Vision and Image Understanding,* vol. 114, no. 6, pp. 712-722, June 2010.

[14] A. Oliva and A. Torralba, “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope,” *International Journal of Computer Vision,* vol. 42, no. 3, pp. 145-175, May 2001.

[15] J. Yao, S. Fidler and R. Urtasun, “Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 702-709, 16-21 June 2012.

[16] D. Munoz, J. Bagnell, N. Vandapel and M. Hebert, “Contextual classification with functional Max-Margin Markov Networks,” *IEEE Conference on Computer Vision and Pattern Recognition,* pp. 975-982, 20-25 June 2009.

[17] E. Sudderth, A. Torralba, W. Freeman and A. Willsky, “Learning hierarchical models of scenes, objects, and parts,” *Tenth IEEE International Conference on Computer Vision,* vol. 2, pp. 1331-1338, 17-21 October 2005.

[18] S. Gould, R. Fulton and D. Koller, “Decomposing a scene into geometric and semantically consistent regions,” *12th International Conference on Computer Vision,* pp. 1-8, 29-02 September - October 2009.

[19] C. Wojek, S. Walk, S. Roth, K. Schindler and B. Schiele, “Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes,” *IEEE Transactions on attern*

Analysis and Machine Intelligence, vol. 35, no. 4, pp. 882 - 897, April 2013.

[20] C. Li, A. Kowdle, A. Saxena and T. Chen, “Toward Holistic Scene Understanding: Feedback Enabled Cascaded Classification Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 7, pp. 1394-1408, July 2012.

[21] S. Ross, D. Munoz, M. Hebert and J. Bagnell, “Learning message-passing inference machines for structured prediction,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2737-2744, 20-25 June 2011.

[22] O. Miksik, D. Munoz, J. Bagnell und M. Hebert, „Efficient temporal consistency for streaming video scene analysis,“ *IEEE International Conference on Robotics and Automation (ICRA),* pp. 133-139, 6-10 May 2013.

[23] A. Geiger, M. Lauer and R. Urtasun, “A Generative Model for 3D Urban Scene Understanding from Movable Platforms,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1945-1952, 20-25 June 2011.

[24] A. Polychronopoulos und A. Amditis, „Revisiting JDL model for automotive safety applications: the PF2 functional model,“ in *IEEE International Conference on Information Fusion*, Florence, 2006.

[25] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” *EEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 2, pp. 807-814, 20-25 June 2005.

[26] A. Elfes, „Using occupancy grids for mobile robot perception and navigation,“ *Computer,* Bd. 22, Nr. 6, pp. 46-57, June 1989.

[27] S. Thrun, „Robotic mapping: A survey,“ in *Exploring Artificial Intelligence in the New Millenium*.

[28] H. Moravec, „Sensor Fusion in Certainty Grids for Mobile Robots,“ *AI Mag.,* Bd. 9, Nr. 2, pp. 61-74, #jul# 1988.

[29] P. Besl und N. D. McKay, „A method for registration of 3-D shapes,“ *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* Bd. 14, Nr. 2, pp. 239-256, Feb 1992.

[30] F. Lu und E. Milios, „Robot pose estimation in unknown environments by matching 2D range scans,“ in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 1994.

[31] A. D. Sappa, R. Herrero, F. Dornaika, D. Gerónimo and A. López, “Road Approximation in Euclidean and v-Disparity Space: A Comparative Study,” *11th International Conference on Computer Aided Systems Theory,* vol. 4739, 1105-1112 17-20 2007.

[32] R. Labayrade, D. Aubert and J.-P. Tarel, “Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation,” *IEEE Intelligent Vehicle Symposium,* vol. 2, pp. 646-651, 12-20 June 2002.

[33] K. Kohara, N. Suganuma, T. Negishi and T. Nanri, “Obstacle Detection Based on Occupancy Grid Maps Using Stereovision System,” *International Journal of Intelligent Transportation Systems Research,* vol. 8, no. 2, pp. 85-95, May 2010.

[34] J. Zhao, J. Katupitiya und J. Ward, *IEEE International Conference on Robotics and Automation,* pp. 529-534, 10-14 April 2007.

[35] D. Pfeiffer and U. Franke, “Efficient representation of traffic scenes by means of dynamic stixels,” *IEEE Intelligent Vehicles Symposium (IV),* pp. 217-224, 21-24 June 2010.

[36] F. Oniga, S. Nedevschi, M. Meinecke and T. B. To, “Road Surface and Obstacle Detection Based on Elevation Maps from Dense Stereo,” *IEEE Intelligent Transportation Systems Conference,* pp. 859-865, 30-03 September - October 2007.

[37] J. Sudano, „Equivalence between belief theories and naive bayesian fusion for systems with independent evidential data: part I, the theory,“ in *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, 2003.

[38] Z. Sun, G. Bebis and R. Miller, “On-road vehicle detection: a review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 5, pp. 694-711, May 2006.

[39] M. Enzweiler, M. Hummel, D. Pfeiffer and U. Franke, “Efficient Stixel-based object recognition,” *IEEE Intelligent Vehicles Symposium (IV),* pp. 1066-1071, 3-7 June 2012.

[40] T. Kowsari, S. Beauchemin and J. Cho, “Real-time vehicle detection and tracking using stereo vision and multi-view AdaBoost,” *14th International IEEE Conference on Intelligent Transportation Systems (ITSC,* pp. 1255-1260, 5-7 October 2011.

[41] C. Wang, H. Zhao, F. Davoine and H. Zha, “A system of automated training sample generation for visual-based car detection,” *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* pp. 4169-4176, 7-12 October 2012.

[42] C.-H. Kuo and R. Nevatia, “Robust multi-view car detection using unsupervised sub-categorization,” *Workshop on Applications of Computer Vision (WACV),* pp. 1-8, 7-8 December 2009.

[43] H. Bay, T. Tuytelaars and L. Van Gool, “SURF: Speeded Up Robust Features,” *Computer Vision – ECCV,* vol. 3951, pp. 404-417, 7-13 May 2006.

[44] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, “Visual categorization with bags of keypoints,” *orkshop on statistical learning in computer vision, ECCV,* 11 May

2004.

[45] P. Dollar, C. Wojek, B. Schiele and P. Perona, “Pedestrian Detection: An Evaluation of the State of the Art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 4, pp. 743-761, April 2012.

[46] R. Benenson, M. Mathias, R. Timofte and L. Van Gool, “Pedestrian detection at 100 frames per second,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2903-2910, 16-20 June 2012.

[47] M. Enzweiler and D. Gavrila, “A Multilevel Mixture-of-Experts Framework for Pedestrian Classification,” *IEEE Transactions on Image Processing,* vol. 20, no. 10, pp. 2967-2979, October 2011.

[48] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* vol. 1, pp. 886-893, 25 June 2005.

[49] X. Wang, T. Han and S. Yan, “An HOG-LBP human detector with partial occlusion handling,” *IEEE 12th International Conference on Computer Vision,* pp. 32-39, 29-02 September - October 2009.

[50] S. Walk, N. Majer, K. Schindler and B. Schiele, “New features and insights for pedestrian detection,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1030-1037, 13-18 June 2010.

[51] Z. Lin und L. S. Davis, „A Pose-Invariant Descriptor for Human Detection and Segmentation,“ *Computer Vision – ECCV,* Nr. 5305, pp. 423-436, 2008.

[52] D. Koller, N. Friedman, L. Getoor and B. Taskar, “Graphical Models in a Nutshell,” in *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, Eds., MIT Press, 2007, pp. 13-55.

[53] E. B. Sudderth, Graphical Models for Visual Object Recognition and Tracking, PhD thesis, Massachusetts Institute of Technology, 2006.

[54] J. Spehr, D. Rosebrock, D. Mossau, R. Auer, S. Brosig and F. Wahl, “Hierarchical scene understanding for intelligent vehicles,” *IEEE Intelligent Vehicles Symposium (IV),* pp. 1142-1147, 5-9 June 2011.

[55] “FASCar,” DLR, Institüt für Verkehrssystemtechnik, [Online]. Available: http://www.dlr.de/fs/desktopdefault.aspx/tabid-1236/1690_read-13097/. [Accessed 24 March 2014].