BAYESIAN SCENE SEGMENTATION INCORPORATING MOTION CONSTRAINTS AND CATEGORY-SPECIFIC INFORMATION

Alexander Bachmann and Irina Lulcheva
Department for Measurement and Control, University of Karlsruhe (TH), 76 131 Karlsruhe, Germany

Keywords: Stereo vision, Motion segmentation, Markov Random fields, Object classification, Global image context.

Abstract: In this paper we address the problem of detecting objects form a moving camera by jointly considering low-level image features and high-level object information. The proposed method partitions an image sequence into independently moving regions with similar 3-dimensional (3D) motion and distance to the observer. In the recognition stage category-specific information is integrated into the partitioning process. An object category is represented by a set of descriptors expressing the local appearance of salient object parts. To account for the geometric relationships among object parts a structural prior over part configurations is designed. This prior structure expresses the spatial dependencies of object parts observed in a training data set. To achieve global consistency in the recognition process, information about the scene is extracted from the entire image based on a set of global image features. These features are used to predict the scene context of the image from which characteristic spatial distributions and properties of an object category are derived. The scene context helps to resolve local ambiguities and achieves locally and globally consistent image segmentation. Our expectations on spatial continuity of objects are expressed in a Markov Random Field (MRF) model. Segmentation results are presented based on real image sequences.

1 INTRODUCTION

One of the cornerstones in the development of automotive driver assistance systems is the comprehensive perception and understanding of the environment in the vicinity of the vehicle. Especially for applications in the road traffic domain the robust and reliable detection of close-by traffic participants is of major interest. In this context, vision sensors provide a rich and versatile source of information (Sivak, 1996), (Rockwell, 1972). Visual object detectors are expected to cope with a wide range of intra-class characteristics, i.e. variations in the visual appearance of an object due to changes in orientation, lighting conditions, scale, etc.. At the same time, these methods must retain enough specificity to yield a minimum amount of misclassifications. Here, most of the approaches developed in the last decades can be partitioned into either: (i) methods based on classification which constrain the detection process to a very specific representation of an object learned from a reference data set or (ii) methods performing object detection by employing local object characteristics on a low level of abstraction using image-based criteria to describe coherent groups of image points as e.g. grey level similarity, texture or motion uniformity of image regions. A major drawback of these methods is the fact that the grouping criteria mostly ignore object-specific properties with the consequence of misdetection rates in cluttered real world scenes that are still prohibitive for most driver assistance applications. This limitation can be weakened by classification methods that have proven to detect a large portion of typical objects at moderate computational cost.

In our approach object detection is performed based on the *relative motion* of textured objects and the observer. The expectation of spatial compactness for most real world objects is expressed by its *position* relative to the observer. To obtain a dense representation of the observed scene, object detection is formulated as an image segmentation task. Here, each image point is tested for consistency with a set of possible hypothesis, each defined by a 3D motion and position. The set of object parameters that best explains the measured quantities of the image point is assigned to the image point.

To further increase the quality of the segmenta-

Bachmann A. and Lulcheva I. (2009).
BAYESIAN SCENE SEGMENTATION INCORPORATING MOTION CONSTRAINTS AND CATEGORY-SPECIFIC INFORMATION.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 291-298
DOI: 10.5220/0001653302910298

tion result, we incorporate information about the objects to be recognised by the system. The integration of object-specific information for driving image segmentation methods has recently developed into a field of active research and seems to be a promising way to incorporate more information into existing low-level object detection methods, see e.g. (Ohm and Ma, 1997; Burl et al., 1998). Our work is inspired by recent research results in human vision, as e.g. (Rentschler et al., 2004), indicating that the recognition and segmentation of a scene is a heavily interweaved process in human perception. Following this biological model our segmentation method is based on low-level features but guided and supported by category-specific information. The question of how to describe this knowledge is very challenging because there is no formal definition of what constitutes an object category. Though most people agree on the choice of a certain object category, there is still much discussion on the choice of an appropriate object descriptor. In our approach the high-level information comprises the appearance of a set of characteristic object parts and its arrangement relative to each other and in the scene. Though good for modeling local object information it fails to capture global consistency in the recognition process, as e.g. the detection of a car in a tree high above the road. We establish global consistency by exploiting the close relationships of certain object categories to the scene of the image. The method characterises a scene by global image features and derives the predicted category likelihood and distribution of an object for a particular scene. We argue that the incorporation of category-specific scene context into our scene segmentation framework can drastically improve the process as (i) insufficient intrinsic object information can be augmented with and (ii) local ambiguities can be better resolved from a global perspective. Figure 1 shows the principle of our probabilistic image segmentation framework.

The remainder of the paper is organised as follows. Section 2 recalls some of the theoretical background that is needed to understand image segmentation as presented here. It is shown how object-specific information can be incorporated into the existing probabilistic framework by means of a sparse object model and category-specific scene information. Section 3 presents the experimental results before conclusions are drawn in Section 4.

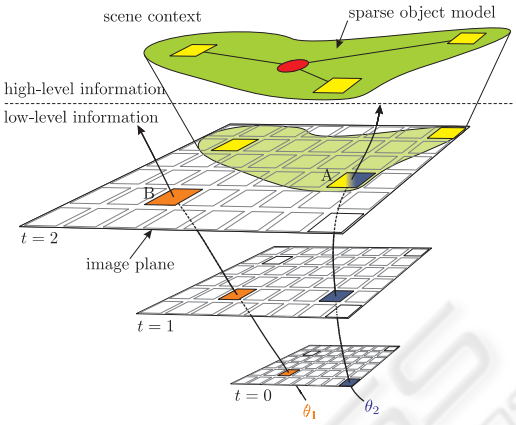


Figure 1: Principle of the combined segmentation process. Image segmentation is performed by a Bayesian maximum a posteriori estimator assigning the most probable object hypothesis to each image point. In the example, image points are assigned to either object hypothesis 1 (B, expressed by θ_1) or object hypothesis 2 (A, expressed by θ_2).

2 SCENE SEGMENTATION USING MRFs

This section outlines the mechanism that evaluates the local and global properties of image points and separates the image accordingly. Notably there are two issues to be addressed in this task: (i) how to encourage the segmentation to consider local properties in the image on a low abstraction level and (ii) how to enforce the process to incorporate category-specific information into the segregation of the image.

First, a number of constraints are formulated that specify an acceptable solution to the problem. In computer vision, commonly the data and prior knowledge are used as constraints. The data constraint restricts a desired solution to be close to the observed data and the prior constraint confines the desired solution to have a form agreeable with the *a priori* knowledge.

The challenging task is the estimation of object parameters $\theta = \{\theta^1, .., \theta^K\}$ given an observation set $\mathbf{Y}$ which has been generated by an unknown and constantly changing number of objects K. Within our framework we solved this by formally expressing the scene segmentation process as *labeling problem*: Let a set of sites (or units) $\mathrm{P} = \{\mathrm{p}_1, .., \mathrm{p}_N\}$, $\mathrm{p}_i \in \mathbb{R}^2$ and a set of possible labels $\mathbf{L}$ be given with one label l_i for each site p_i specifying the process which generated the data. l_i is a binary vector such that $\mathrm{l}_i^j = 1$ if object j generated the data at site p_i. The desired labeling is then a mapping $\mathbf{l} : \mathrm{P} \mapsto \mathbf{L}$ that assigns a unique la-

bel to each site. The labeling $\mathbf{l} = (l(p_1),..,l(p_N)) = (l_1,..,l_N)$ shall ascertain that (i) the data of all sites with identical label exhibit similarity w.r.t. some measure and that (ii) the labeling conforms with the *a priori* knowledge.

Taking a Bayesian perspective the posterior probability of a labeling $\mathbf{l}$ can be formulated

$$P(\mathbf{l}|\mathbf{Y},\theta) = \frac{P(\mathbf{Y}|\mathbf{l},\theta)P(\mathbf{l}|\theta)}{P(\mathbf{Y},\theta)}\,, \tag{1}$$

where we try to find the labeling $\mathbf{l}$ which maximises $P(\mathbf{l}|\mathbf{Y},\theta)$. Here, $P(\mathbf{Y}|\mathbf{l},\theta)$ states the data constraint parametrised by object parameter vector θ. $P(\mathbf{l}|\theta)$ states the prior term. Obviously, $P(\mathbf{Y})$ in Equation (1) does not depend on the labeling $\mathbf{l}$ and can thus be discarded during the maximization. By rearranging Equation (1) the *maximum a posteriori* (MAP) estimate of a labeling can be expressed

$$\widehat{\mathbf{l}} = \arg\max_{\mathbf{l}} \underbrace{P(\mathbf{Y}|\mathbf{l},\theta)}_{\text{data term}} \underbrace{P(\mathbf{l}|\theta)}_{\text{prior term}}\,. \tag{2}$$

Assuming the observations $\mathbf{Y}$ to be i.i.d. normal, the first term in Equation (2) can be written

$$\begin{aligned} P(\mathbf{Y}|\mathbf{l},\theta) &= \prod_{i=1}^{N} P(y_i|l_i,\theta) \\ &\propto \prod_{i=1}^{N} \exp(-E_i(y_i|l_i,\theta)). \end{aligned} \tag{3}$$

E_i denotes an energy functional, rating observation y_i given label vector l_i and object parameter vector θ.

If there exists no prior knowledge about the values of θ (i.e. $P(\theta) = \text{const.}$) prior expectations on $\mathbf{l}$ can be modelled using MRFs. An MRF is defined by the property $P(l_i\,|\,l_1,..,l_{i-1},l_{i+1},l_N) = P(l_i\,|\,l_j,\forall_j \in \mathcal{G}_i)$, with $\mathcal{G}_i$ being the neighbourhood set of image point p_i. The system must fulfill the constraints (i) $p_i \notin \mathcal{G}_i \forall l_i$, no site is its own neighbour and (ii) $p_i \in \mathcal{G}_j \Leftrightarrow p_j \in \mathcal{G}_i$, if p_i is a neighbour of p_j, then p_j is also a neighbour of p_i. Due to the equivalence of MRF and Gibbs distributions, see e.g. (Besag, 1974), an MRF may be written as $P(\mathbf{l}) = \frac{1}{\mathbf{Z}}\exp(-V_k(\mathbf{l}))$, where $V_k(\mathbf{l}) \in \mathbb{R}$ is referred to as clique potential which only depends on those labels of $\mathbf{l}$ whose sites are elements of clique k. A clique $k \subseteq P$ is any set of sites such that any of its pairs are neighbours. We model the clique potential for 2-element cliques $k = \{p_i, p_j\}$ with $|p_i - p_j| = 1$ using an extension of the generalised Potts model ((Geman and Geman, 1984))

$$V_{\{i,j\}}(l_i,l_j) = \begin{cases} \lambda & if\ l_i \neq l_j \\ 0 & \text{otherwise} \end{cases}\,, \tag{4}$$

favoring identical labels at neighbouring sites. The coefficient λ modulates the effect of the prior term and therefore the degree of label smoothness in the segmentation result. The generalised Potts model is a natural and simple way to define a clique potential function that describes the smoothness of neighbouring points. With Equation (3)–(4), Equation (2) evolves to

$$\begin{aligned} \widehat{\mathbf{l}} &= \arg\min_{\mathbf{l}} \sum_{i=1}^{N} E_i(y_i|l_i,\theta) + \sum_{i=1}^{N}\sum_{j\in\mathbf{k}} V_{\{i,j\}}(l_i,l_j) \\ &= \arg\min_{\mathbf{l}} \Psi(E,\mathbf{l},\theta)\,. \end{aligned} \tag{5}$$

2.1 Low-level Information

Concerning the data term, in (Bachmann and Dang, 2008) excellent results have been achieved using the property *object motion*. Here, objects are specified by a 6-degree-of-freedom (dof) parametric motion model representing the motion of an image region by parameter vector $\mathbf{v}$. The similarity between expected and observed object motion is expressed by evaluating the similarity between expected image texture $G_{t-1}(p_i;\mathbf{r};\mathbf{v})$ derived from motion profile $\mathbf{v}$ and observed image texture $G_t(p_i;\mathbf{r})$ within a block of size B centered around image point p_i

$$E_i(\varepsilon_i^{\mathbf{v}}|l_i,\mathbf{v}) = \sum_{\mathbf{r}\in B} \left(G_t(p_i;\mathbf{r}) - G_{t-1}(p_i;\mathbf{r};\mathbf{v})\right)^2\,, \tag{6}$$

with $\varepsilon_i^{\mathbf{v}}$ stating the residual at image position p_i.

This object model has been further extended by the object position ξ relative to the own vehicle, i.e. $\theta = (\mathbf{v},\xi)$. The relative object position is expressed by the mean disparity ξ^{Δ} of all image points assigned to the respective label and the assignment energy of image point p_i given an object label is

$$E_i(\varepsilon_i^{\xi}|l_i,\xi) = \frac{\left(\Delta_i - \xi^{\Delta}\right)^2}{2\sigma_{\xi^{\Delta}}^2}. \tag{7}$$

$\sigma_{\xi^{\Delta}}$ states the extension of the object in terms of the variance of the disparity values assigned to the object label.

The object model presented above allows to segregate an image sequence into K distinct regions with each region being defined as homogenously moving object at a certain distance to the observer, i.e. $\mathbf{L} = \{\texttt{background}, \texttt{object 2}, .., \texttt{object K}\}$, with image regions moving static relative to the observer (as e.g. trees, buildings, etc.) being labeled $\{\texttt{background}\}$.

With the intention to classify every image point into a meaningful semantic category and due to the well-known limitations of motion-based segmentation methods (as e.g. the aperture problem or poorly textured image regions) the next step is to incorporate category-specific information into the segmentation process.

2.2 Category-specific Information

Therefore we extend our algorithm to perform interleaved object recognition and segmentation. To achieve this, the object parameter vector θ is extended by model parameter Φ expressing the configuration of an object of a certain category c_O. An image point is either assigned to one of the defined object categories $\{\texttt{car, bicycle, pedestrian}\} \in c_O$ or, if none of the categories adequately describes the image point, $\{\texttt{obstacle}\} \in c_O$. To incorporate object categories into our segmentation scheme, Equation (5) is extended to

$$\Psi(E,\mathbf{l},\theta) = \underbrace{\Psi(E,\mathbf{l},\mathbf{v},\xi)}_{\text{object motion \& position}} + \underbrace{\Psi(E,\mathbf{l},\Phi)}_{\text{object category}}, \quad (8)$$

with

$$\Psi(E,\mathbf{l},\Phi) = \sum_{i=1}^{N} E_i(\varepsilon_i^{\Phi}|\mathrm{l}_i,\Phi). \quad (9)$$

The function $E(\varepsilon_i^{\Phi}|\mathrm{l}_i,\Phi)$ ascertains that image points falling close to a given object description would more likely carry the object category label and vice versa. The energy functional has the form

$$E_i(\varepsilon_i^{\Phi}|\mathrm{l}_i,\Phi) = -\log P(\varepsilon_i^{\Phi}|\mathrm{l}_i,\Phi). \quad (10)$$

For this work $P(\varepsilon_i^{\Phi}|\mathrm{l}_i,\Phi)$ is defined as

$$P(\varepsilon_i^{\Phi}|\mathrm{l}_i^j = 1,\Phi) = \frac{1}{1+d(\mathrm{p}_i,\Phi^j)}, \quad (11)$$

with $d(\mathrm{p}_i,\Phi^j)$ expressing the distance from image point p_i to the object that is parametrised by Φ^j.

In this work, an object of a certain category is characterised by the local appearance of a set of n salient parts $\Phi = (\phi_1,...,\phi_n)$, with $\phi_i = (x_i,y_i,z_i,\rho_i)$ stating the location of the i-th part in 3D space and ρ_i being the scale factor. Depth z_i is obtained from a calibrated stereo camera setup (Dang et al., 2006). The structural arrangement of the parts comprising an object is expressed by the spatial configuration of Φ. Spatial relationships between parts in the sparse object model are captured by parameter $\mathbf{s}$. The local appearance of each part is characterised by parameter $\mathbf{a}$. The pair $\mathbf{M} = (\mathbf{s},\mathbf{a})$ parameterises an object category. Again, using Bayes rule the probability of an object being at a particular location, given fixed model parameters, can be written

$$P_{\mathbf{M}}(\Phi|\mathbf{Y}) \propto P_{\mathbf{M}}(\mathbf{Y}|\Phi)P_{\mathbf{M}}(\Phi). \quad (12)$$

Above, $P_{\mathbf{M}}(\mathbf{Y}|\Phi)$ is the likelihood of the feature points depicting an object for a certain configuration of the object parts. The second term in Equation (12) is the prior probability that the object obeys the spatial configuration Φ. Assuming the object is present in an image, the location that is most likely its true position is the one with maximum posterior probability

$$\hat{\Phi} \propto \arg\max_{\Phi} P_{\mathbf{M}}(\mathbf{Y}|\Phi)P_{\mathbf{M}}(\Phi). \quad (13)$$

Local Appearance. The image evidence $P_{\mathbf{M}}(\mathbf{Y}|\Phi)$ of the individual parts in the sparse object model is modelled by its local appearance. The part appearance $\mathbf{a}_i$, characterising the i-th part of a certain object model is extracted from an image patch centered on $\Pi(\phi_i)$, where $\Pi(\cdot)$ symbolises the projection of a scene point onto the image plane. The object-characteristic appearance of each image patch $i \in (1,..,n)$ has been learned from a set of labeled training images. In this work three types of appearance measures $\mathbf{a}_i = \{a_i^1; a_i^2; a_i^3\}$ have been used to describe an object:

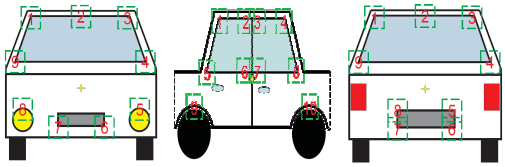
- **Texture information** a_i^1, the magnitude of each pixel within the patch is stacked into a histogramm vector to express the texture.
- **Shape information** a_i^2, the Euclidean distance transform of the edge map within the patch expresses the shape.
- **Height information** a_i^3, the characteristic height of ϕ_i above the estimated road plane expresses the relative location in the scene.

The resulting patch responses constitute a vector of local identifiers for each object category. The model parameters have been learned from a set of labeled training images in order to generate a representative description of the local appearance of an object category. Prominent regions have been extracted from the image using the Harris interest point detector (Harris and Stephens, 1988) and a corner detector based on curvature scale space technique as described in (He and Yung, 2004). For object part ϕ_i and observation vector $\mathbf{Y}$ follows the model likelihood

$$P_{\mathbf{M}}(\mathbf{Y}|\Phi) = \prod_{i=1}^{n} P_{\mathbf{M}}(\mathbf{Y}|\phi_i). \quad (14)$$

The likelihood function measures the probability of observing $\mathbf{Y}$ in an image, given a particular configuration Φ. Intuitively, the likelihood should be high when the appearance of the parts agree with the image data at the positions they are placed, and low otherwise. Figure 2 shows the sparse object model of object category `car`.

Structural Prior. What remains is to encode the assumed spatial relationships among object parts. As



(a) (b) (c)

Figure 2: Sparse representation of object category `car`: **(a)** front view, **(b)** side view, **(c)** rear view. The parts used in the training stage are marked with green rectangles containing the part-ID.

presented in (Bachmann and Dang, 2008) the assumption can be made that the part locations are independent

$$P_{\mathbf{M}}(\Phi) = \prod_{i=1}^{n} P_{\mathbf{M}}(\phi_i)\,. \tag{15}$$

Here, only the metric height above the estimated road plane has been used as structural information. Maximizing $P_{\mathbf{M}}(\Phi|\mathbf{Y})$ is particularly easy as $P_{\mathbf{M}}(\mathbf{Y}|\Phi)P_{\mathbf{M}}(\Phi)$ can be solved independently for each ϕ_i. For n parts and N possible locations in the image this can be done in $O(nN)$ time. A major drawback of this method is that it encodes only weak spatial information and is unable to accurately represent objects composed of various parts.

The most obvious approach to represent multi-part objects is to make no independence assumption on the locations of different parts. Though theoretically appealing the question of how to efficiently perform inference on this spatial prior is not trivial.

A balance between the inadequate independence assumption and the strong but hard to implement full dependency between object parts is assumed by maintaining certain conditional independence assumptions. These assumptions can be elegantly represented using an MRF where the location of part ϕ_i is independent of the values of all other parts $\phi_j\,, j \neq i$, conditioned on the values of the neighbours $\mathcal{G}_i$ of ϕ_i in an undirected graph $G(\Phi, E)$. The structural prior is characterised by pairwise only dependencies between parts.

Sparse Object Model. The spatial prior is modeled as a star structured graph with the location of the object parts being conditioned on the location of reference point ϕ_R. For a better understanding ϕ_R can be interpreted as center of mass of the object. All object parts arranged around ϕ_R are independent of one another. A similar model is also used by e.g. (Crandall and Huttenlocher, 2007; Fischler and Elschlager, 1973). Let $G = (\Phi, E)$ be a star graph with central node ϕ_R. Graphical models with a star structure have a straight forward interpretation in terms of the conditional distribution

$$P_{\mathbf{M}}(\Phi) = P(\phi_R) \prod_{i=1}^{n} P_{\mathbf{M}}(\phi_i|\phi_R)\,. \tag{16}$$

Reference point ϕ_R acts as the anchor point for all neighbouring parts. The positions of all other parts in the model are evaluated relative to the position of this reference point. In this work we chose ϕ_R to be virtual, i.e. there exists no measurable quantity that indicates the existence of the reference point itself. We argue that this makes the model insensitive to partial object occlusion and, therefore, to the absence of reference points. $P_{\mathbf{M}}(\Phi)$ is modelled using a Mixture of Gaussian (MoG). The model parameter subset $\mathbf{M} = (\mathbf{s}, \cdot)$, with mean $\mu_{i,R}$ and covariance $\sigma_{i,R}$ stating the location of ϕ_i relative to the reference point ϕ_R, has been determined in a training stage.

An optimal object part configuration (see Equation (13)) can be written in terms of observing an object at a particular spatial configuration $\Phi = (\phi_1, .., \phi_n)$, given the observations $\mathbf{Y}$ in the image. With the likelihood function of seing object part i at position ϕ_i (given by Equation (14)) and the structural prior in Equation (16) this can be formulated as

$$P_{\mathbf{M}}(\Phi|\mathbf{Y}) \propto P(\phi_R)\Gamma(\phi_R|\mathbf{Y})\,, \tag{17}$$

where the quality of the reference point ϕ_R relative to all parts ϕ_i within the object definition is written

$$\Gamma(\phi_R|\mathbf{Y}) = \max_{\phi} \prod_{i=1}^{n} P_{\mathbf{M}}(\phi_i|\phi_R) P_{\mathbf{M}}(\mathbf{Y}|\phi_i)\,. \tag{18}$$

What we are interested in, is finding the best configuration for all n parts of the object model relative to ϕ_R. To reduce computational costs only points are further processed with a likelihood $P_{\mathbf{M}}(\mathbf{Y}|\phi_i) > T$, where T is the acceptance threshold for the object hypothesis to be true. This results in a number of candidates m for each object part i. As this is computationally infeasible $(O(m^n))$ for large growing n we propose a greedy search algorithm to maximise $P_{\mathbf{M}}(\Phi|\mathbf{Y})$ over all possible configurations $\{\phi_i^j : i = 1, .., n; j = 1, .., m\}$ as outlined in Table1.

2.3 Context Information

The MRF presented above efficiently models local image information consisting of low-level features enriched by high-level category-specific information.

However, context information capturing the overall global consistency of the segmentation result has been ignored so far. By introducing a set of semantic categories into the segmentation process, it is now possible to derive category-specific object characteristics not only on a local, object-intrinsic level but

Table 1: Iterative search algorithm.

1. compute candidates $\phi^j, j = (1,..,m)$ for which $P_{\mathbf{M}}(\mathbf{Y}|\phi_i^j) > T$
2. initialise $\phi_i^j, j \in (1,..,m)$ for object part $i = 1$; set $\mathbf{k} = i$; for each candidate ϕ^j...
   (a) ...vote for reference point ϕ_R^j based on part location ϕ_i
   (b) ...set $i = i+1$ and $\mathbf{k} = [\mathbf{k}; i]$
   (c) ...back-project ϕ_i from ϕ_R^j and compute $P_{\mathbf{M}}(\Phi^*|\mathbf{Y})$, with $\Phi^* = (\phi_{\mathbf{k}})$
   (d) ...IF $P_{\mathbf{M}}(\Phi^*|\mathbf{Y}) > T$: go back to (a); ...ELSE: end

also on a global scale, expressing the relationships between labels and global image features. In this work this is the predicted distribution of object categories in the image which helps to achieve globally consistent recognition. Based on the work presented in (Bachmann and Balthasar, 2008) we exploit the relation between the expected distribution of a certain category and the scenery. The scene-based information is formally introduced into our framework by extending Equation (5) with a context-aware object prior predicting the distribution of category labels

$$\Psi(E,\mathbf{l},\theta) = \underbrace{\Psi(E,\mathbf{l},\mathbf{v},\xi,\Phi)}_{\text{local information}} + \underbrace{\sum_{i=1}^{N} G_i(\mathrm{l}_i|\mathbf{Y})}_{\text{context information}} , \quad (19)$$

with category context potential

$$G_i(\mathrm{l}_i|\mathbf{Y}) = \log P(\mathrm{l}_i|\mathbf{M}_C) . \quad (20)$$

$G_i(\cdot)$ predicts the label l_i from a global perspective using global image features $\mathbf{M}_C$. The global features characterise the entire image in terms of magnitude and orientation of edges in different image resolutions. For this work we defined a set of scene categories $c_S = \{$`open, semi-open, closed`$\}$, each expressed by a unique feature vector $\mathbf{M}_C$, describing the *openness* of the scene. This information is used to derive the distribution of category labels and category probability in the image. The feature vector $\mathbf{M}_C$ for each specific scene has been calculated from a training data set and is formally expressed by a mixture of Gaussian model. The relationships between the contextual features and a specific object category c_O has been learned in a training stage as presented in (Bachmann and Balthasar, 2008). Given an input image, the prior probability of an object category c_O is expressed as its marginal distribution over all scene categories c_S whereas the scene similarity of the input image (expressed by $\mathbf{M}_C^{\text{obs}}$) to the defined scene categories is determined by calculating the joint probability with the single components in $\mathbf{M}_C$.

3 EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the object detection approach developed in the previous sections. The results are based on image sequences of typical urban traffic scenarios. The algorithm is initialised automatically by scanning for the actual number of dominant motions in the scene. Concerning the motion of the observer, the road plane is determined at the beginning of the image sequence as described in (Duchow et al., 2006). Thus, the motion profile of the observer can be determined by sampling feature points exclusively from the region that is labeled as road plane and therefore static relative to the observer. During the segmentation process, the motion profiles are refined and updated continuously with the motion tracker scheme described in (Bachmann and Dang, 2008). Regarding the relative importance of data and smoothness term in the segmentation process, the regularization factor was adapted empirically to values between $\lambda = (0.05,..,0.5)$.

The confidence of an image point to be part of an object hypothesis, i.e. label, is calculated based on its relative motion, its position and similarity to the defined object categories. The image point is assigned to the label with highest confidence. The training data for object category `car` as presented here was extracted from an image data base of 160 images.

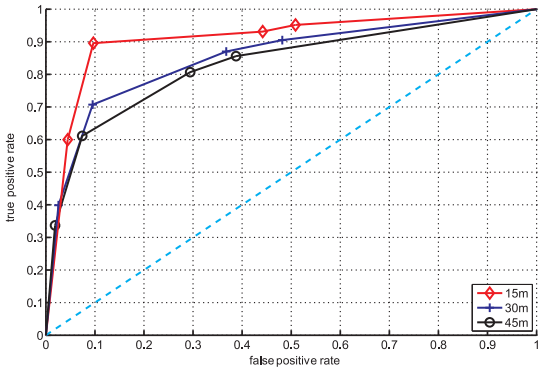


Figure 3: ROC-curve for rear view of object category `car` as a function of the distance from the observer.

Figure 3 shows that a threshold value of $T \approx 0.6$ yields a good compromise between a reasonable true positive rate and a false positive rate at relative low values.

Figure 4 shows some of the detection results for object category `car`. The model was learned from

labeled data. The patch-size of the extracted interest points was scale-normalised based on a predefined reference scale.

Figure 4: Detection results (threshold value $T = 0.5$) for object category `car`. The figure on the bottom right shows a false positive.

Figure 5 shows the classification results for detected objects solely based on scene context. No local category information has been integrated. It can be seen that the integration of global information is useful as a first process of image recognition. Joint use of the proposed local object detector together with category-specific scene context improves the recognition accuracy as shown in Figure 6.

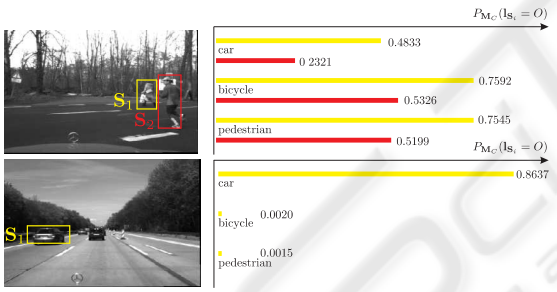


Figure 5: **Left:** Detected objects based on the local object properties *motion similarity* and *position*. **Right:** Prior probability $P_{\mathbf{M}_C}(\mathbf{l}_{\mathbf{S}_i} = O)$ of image region $\mathbf{S}_i$ to belong to object category $O = \{$`car, bicycle, pedestrian`$\}$ solely based on scene context information.

Here, the object detection and segmentation results for different traffic scenes is depicted. As in this work only object category `car` is known locally to the system, detected objects that are labeled `bicyclist` or `pedestrian` rely solely on the global context of the image. In the initialization phase of the segmentation process the motion estimates for the labels are inaccurate. Therefore the segmentation is mainly driven by the appearance-based confidence measure. With increasing accuracy and distinctiveness of the labels motion profile, the influence of the motion cue increases. In most cases, it takes less than 3 frames to partition the image into meaningful regions.

4 CONCLUSIONS

This paper has presented a MRF for pixel-accurate object recognition that models local object information and global information explicitly. The local information consists of a set of distinct 6-dof motion profiles, positions and - on a high abstraction level - the local appearance similarity to the trained object category `car`. Distinctive, local object descriptors and a structural prior on the object-parts configuration have been extracted from a set of sample images. The structural relationships among object parts has been modelled as sparse structural prior. Object recognition is realised by an iterative method that finds an optimal configuration of the object parts based on the local appearance in the image and its spatial arrangement. Global information is derived from scene-based information generated according to the scene of the input image. As the occurrence of object categories is closely related to the scene of the image, scene context is exploited to derive characteristic category distributions and probabilities. It has been shown that the joint use of a local object detector and scene context improves the recognition accuracy. Under the assumption of motion and category homogeneity within the boundaries of an object, spatial consistency has been modelled through a Markov Random Field.

In ongoing work, we expect to increase the performance of the method by further refining and extending the sparse object model description. We suppose to increase the quality of the classification process by making the object appearance descriptors invariant to object orientation and rotation. Additionally, the performance shall be increased by an exhausting training of different object categories. Furthermore it is intended to speed up the search algorithm that incorporates the object spatial configuration to make the entire process computationally feasible.

ACKNOWLEDGEMENTS

This work was partly supported by the Deutsche Forschungsgemeinschaft DFG within the collaborative research center ‘Cognitive Automobiles’.

REFERENCES

Bachmann, A. and Balthasar, M. (2008). Context-aware object priors. In *IEEE IROS 2008; Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)*, Nice, France.

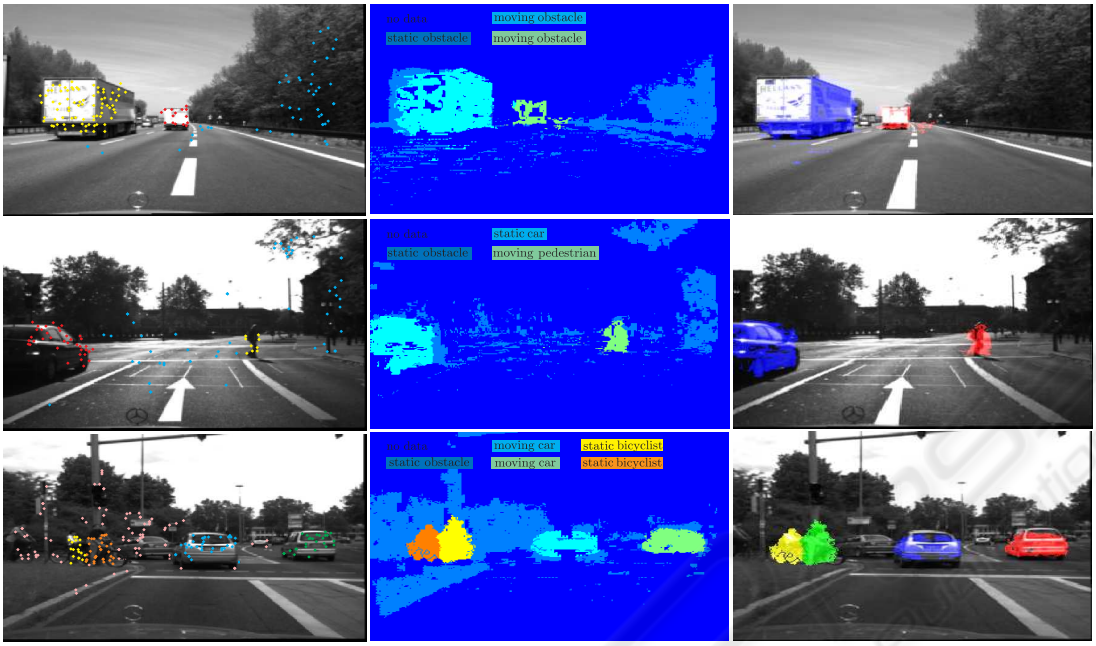


Figure 6: **Left:** Scenes with a variable number of moving objects. The 6–dof motion for each object is determined based on a set of interest points extracted from the respective object. Coloured markers indicate the tracked points. **Middle:** The resulting segmentation map. Each image point is assigned the label containing the most probable motion profile and position. **Right:** The segmented image. Image points assigned to an object label are highlighted.

Bachmann, A. and Dang, T. (2008). Improving motion-based object detection by incorporating object-specific knowledge. *International Journal of Intelligent Information and Database Systems (IJIIDS)*, 2(2):258–276.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, Series B 36(2):192–236.

Burl, M. C., Weber, M., and Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. *Lecture Notes in Computer Science*, 1407:628ff.

Crandall, D. and Huttenlocher, D. (2007). Composite models of objects and scenes for category recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR '07*, pages 1–8.

Dang, T., Hoffmann, C., and Stiller, C. (2006). Self-calibration for active automotive stereo vision. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, Tokyo.

Duchow, C., Hummel, B., Bachmann, A., Yang, Z., and Stiller, C. (2006). Akquisition, Repraesentation und Nutzung von Wissen in der Fahrerassistenz. In *Informationsfusion in der Mess- und Regelungstechnik 2006, VDI/VDE-GMA*. Eisenach, Germany.

Fischler, M. and Elschlager, R. (1973). The representation and matching of pictorial structures. *IEEE Trans. Comput.*, 22(1):67–92.

Geman, S. and Geman, D. (1984). Stochstic relaxation, Gibbs distribution, and the Bayesian restoration of images. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, volume 6, pages 721–741.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Fourth Alvey Vision Conference, Manchester*, pages 147–151.

He, X. and Yung, N. (2004). Curvature scale space corner detector with adaptive threshold and dynamic region of support. In *17th International Conference on Pattern Recognition*, volume 2, pages 791–794, Washington, DC, USA. IEEE Computer Society.

Ohm, J.-R. and Ma, P. (1997). Feature-Based cluster segmentation of image sequences. In *ICIP '97-Volume 3*, pages 178–181, Washington, DC, USA. IEEE Computer Society.

Rentschler, I., Juettner, M., Osmana, E., Mueller, A., and Caell, T. (2004). Development of configural 3D object recognition. *Elsevier - Behavioural Brain Research*, 149(149):107–111.

Rockwell, T. (1972). Skills, judgment, and information acquisition in driving. *Human Factors in Highway Traffic Safety Research*, pages 133–164.

Sivak, M. (1996). The information that drivers use: is it indeed 90% visual? *Perception*, 25(9):1081–1089.