2013 IEEE Intelligent Vehicles Symposium (IV)
June 23-26, 2013, Gold Coast, Australia

From Stixels to Objects - A Conditional Random Field based Approach

Friedrich Erbs[1], Beate Schwarz[2] and Uwe Franke[1]

***Abstract*— Detection and tracking of moving traffic participants like vehicles, pedestrians or bicycles from a mobile platform using a stereo camera system plays a key role in traffic scene understanding and for future driver assistance and safety systems. To this end, this work presents a Bayesian segmentation approach based on the Dynamic Stixel World, an efficient super-pixel object representation. The existence and state estimation of an (initially) unknown number of moving objects and the detection of stationary background is formulated as a time-recursive energy minimization problem that can be solved in real-time by means of the alpha-expansion multi-class graph cut optimization scheme. In order to handle noise, this approach integrates 3D and motion features as well as spatio-temporal prior knowledge in a probabilistic conditional random field (CRF) framework. An optional fusion step with an additional radar sensor combines the advantages of both measuring instruments and yields superior overall results. The performance and robustness of the presented approach is evaluated quantitatively in various challenging traffic scenes.**

I. INTRODUCTION

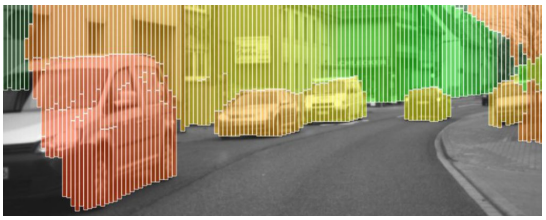
The use of camera systems for capturing dense 3D and motion information has increased tremendously in the last years and has allowed the development of numerous driver assistance systems. In view of the steadily increasing number of driver assistance systems, the introduction of a medium-level representation called Stixel World [23], [22] has proven to be highly advantageous. This is because the Stixel World yields the important freespace information and it provides an efficient object representation that is required by many driver assistance applications. Furthermore, it allows reducing the computational burden for these subsequent applications, in some cases by several orders of magnitude, see e.g. [11], [10], [3].

In this contribution, an Expectation-Maximization-like (EM) [9] CRF approach is derived that can detect and track the moving objects in stereo image sequences based on the Dynamic Stixel World. The actual number of objects is part of the objective function to be minimized. The main steps of the presented approach are summarized in Figure 1. Firstly, a dense depth image is computed. These experiments use the Semi-Global Matching (SGM) algorithm [16], [15], as shown in Figure 1(b). Secondly, the multi-layered Stixel World [23] shown in Figure 1(c) is computed. Subsequently, the stixels are tracked over time to estimate their motion state by applying the 6D-Vision principle [14], [22] as shown in Figure 1(d).

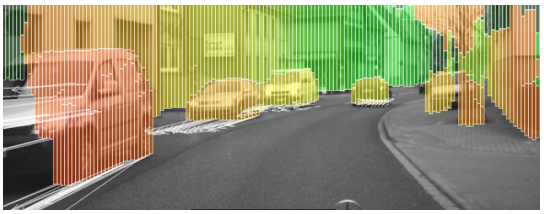
However, the Dynamic Stixel World does not contain any information about their relation to each other. This independence assumption could result in inconsistent scene interpretations in the event of strong noise under adverse conditions.

(a) Left original gray value image.

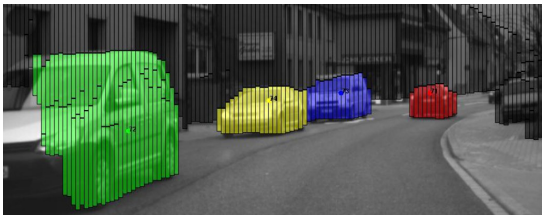
(b) SGM [15] stereo image. The distance is color encoded ranging from red (close) to green (far away).



(c) Multi-Layered Stixel World [21].



(d) Dynamic Stixel World [22]. The arrows show the predicted position of the stixels for the next half second.



(e) Object segmentation result. The color encodes the different object classes, the static background is shown in black.

Fig. 1. Processing chain of the segmentation.

[1]The author is with Daimler AG, Image Understanding, Boeblingen, Germany.

[2]B. Schwarz is with the Department of Mathematics, HFT Stuttgart, 70174 Stuttgart, Germany.

978-1-4673-2754-1/13/$31.00 ©2013 IEEE

For this reason, a grouping of the stixels into objects is desirable. In this work, a segmentation approach for this step is proposed. The segmentation step partitions the Dynamic Stixel World into several moving objects, exemplarily shown in Figure 1(e).

The remainder of this article is structured as follows: Section II briefly reviews related work. Section III derives the presented approach from probability theory and Section IV evaluates this approach in real-world scenes using ground-truth material. Finally, Section V concludes this contribution.

II. RELATED WORK

Segmentation in the presence of strong noise requires strong regularization and redundancies. Especially Conditional Random Fields (CRFs) [19] have proven to be remarkably powerful since they can model spatio-temporal correlations and can be solved nearly optimally by means of efficient inference methods [17]. Classical CRF approaches consider segmentation as a labeling problem, where each (super-) pixel needs to be assigned to exactly one object class like street, sky or car, c.f. [12], [6], [7]. These approaches are restricted to static scenes and do not allow separating between *different instances* of one object class, e.g. between different moving objects.

Accordingly, the CRF concept has been extended for object detection, see e.g. [30], [18]. In [30], the CRF acts as a mediator between the output of different part-based classifier responses ensuring some global ordering relations. However, this approach considers a pure appearance-based classification for static scenes, in contrast to the scope of the present approach.

Besides, for many applications the discrete class choice made in the contributions mentioned above seems imprecise since objects have continuous parameters. Markov Random Fields have been generalized to the continuous domain requiring user interaction as done in [25], or [28]. In [29], Unger et al. proposed a joint segmentation and motion estimation approach in the continuous domain. However, the authors conclude that their approach is still too slow for many practical purposes. Besides that, the inclusion of further segmentation classes during the optimization is done in an ad-hoc manner. The approach presented here derives this quantity based on probability theory.

In [13], Wang et al. propose a vertical extension to the binary graph cut segmentation problem for an unknown number of class labels using different split-and-merge schemes.

An Expectation-Maximization approach related to the one being presented has been published by Bachmann in [1]. However, the regularization on the pixel level was found to be often insufficient in the presence of strong noise. Besides that, the actual number of objects is derived heuristically in that work.

Another concept is layer decomposition like [27] which is still too slow for real-time applicability.

The most similar work has been done by Barth et al. [2], which use dense depth and motion information for multi-class traffic scene segmentation. However, using the Stixel World instead of individual pixels is more robust and is at least three orders of magnitude faster, thus, the Stixel World enables automotive real-time operation in the first place. Besides that, the approach by Barth does not estimate the motion state of the moving objects by itself, this information has to be provided externally.

Finally, in [11] the present authors proposed a related approach also based on the Dynamic Stixel World. However, this approach performs motion class segmentation into a predefined number of classes. In the present work, the actual motion state is arbitrary and part of the estimation process.

III. PROBABILISTIC FORMULATION

In this section, the segmentation task is formulated as a Bayesian optimization problem. The given stereo camera system records an image sequence $\mathcal{I}$ with dense stereo information, that is subsequently segmented into the multi-layered Stixel World as proposed in [21]. This (static) Stixel World partitions an input image $I^t \in \mathcal{I}$ at time step t column-wise into several layers of one of the two classes $\mathcal{C}_{\text{Stixel}} \in \{\text{street, obstacle}\}$, c.f. 1(c). In the following, the street area is left unchanged and the focus is on obstacle stixels. Subsequently, the stixels are tracked over time in order to estimate their motion state [22].

In summary, each stixel with index i is defined by solely five observations. That is its 3D world position $\{X_i^t,\, H_i^t,\, Z_i^t\}$, where H_i^t denotes the height of the stixel relative to the camera coordinate system, X_i^t is the lateral position pointing to the right and Z_i^t is the longitudinal position in driving direction and its velocity $\{\dot{X}_i^t,\, \dot{Z}_i^t\}$. Moving objects such as cars or bicycles are assumed to move in the ground plane, so it is sufficient to estimate a 2D motion vector. These five observations form a feature vector for each stixel, $\vec{z}_i^{\,t} = \{\dot{X}_i^t,\, \dot{Z}_i^t,\, X_i^t,\, H_i^t,\, Z_i^t\}^{\mathrm{T}}$. These feature vectors are again combined in a measurement array $\mathcal{Z}^t = \{\vec{z}_1^{\,t},\, \ldots,\, \vec{z}_N^{\,t}\}$.

Now, let $\mathbf{L}^t = \{l_1^t,\, \ldots, l_N^t\}^{\mathrm{T}}$ denote a labeling for a given input image I^t containing N dynamic stixels. A labeling assigns each stixel to exactly one of K moving objects or to static background, $l_i^t \in \{O_1,\, O_2,\, \ldots O_K,\, \text{bg}\}$.

The most probable labeling maximizes

$$p\left(\mathbf{L}^t \mid \mathcal{Z}^t, \mathbf{L}^{t-1}\right) \propto \underbrace{p\left(\mathcal{Z}^t \mid \mathbf{L}^t\right)}_{\text{Data Term}} \cdot \underbrace{p\left(\mathbf{L}^{t-1} \mid \mathbf{L}^t\right)}_{\text{Temporal Term}} \cdot \underbrace{p\left(\mathbf{L}^t\right)}_{\text{Prior Term}}, \tag{1}$$

taking into account the previous segmentation $\mathbf{L}^{t-1}$. The probability of such a labeling is modeled as a CRF with a maximum clique size of two, i.e.

$$\begin{aligned} p\left(\mathcal{Z}^t \mid \mathbf{L}^t\right) &\propto \prod_{i=1}^{N} p\left(\vec{z}_i^{\,t} \mid l_i^t\right) \cdot \prod_{(i,j)\in\mathcal{N}_2} p\left(\vec{z}_i^{\,t}, \vec{z}_j^{\,t} \mid l_i^t, l_j^t\right), \\ p\left(\mathbf{L}^{t-1} \mid \mathbf{L}^t\right) &\propto \prod_{i=1}^{N} p\left(l_i^{t-1} \mid l_i^t\right), \\ p\left(\mathbf{L}^t\right) &\propto \prod_{i=1}^{N} p\left(l_i^t\right) \cdot \prod_{(i,j)\in\mathcal{N}_2} p\left(l_i^t, l_j^t\right). \end{aligned} \tag{2}$$

In this context, $\mathcal{N}_2$ denotes the set of adjacent stixels. Equivalently to Equation 1 one minimizes $E := -\log p\left(\mathcal{Z}^t \mid \mathbf{L}^t\right)$

$$E \propto \underbrace{-\sum_{i=1}^{N} \log p\left(l_i^t\right)}_{:=\log Q(\mathbf{L}^t)} \underbrace{-\sum_{i=1}^{N} \log p\left(l_i^{t-1} \mid l_i^t\right)}_{:=\log p(\mathbf{L}^{t-1}|\mathbf{L}^t)} - \underbrace{\sum_{i=1}^{N} \log p\left(\vec{z}_i^t \mid l_i^t\right)}_{:=\log Q(\mathcal{Z}^t|\mathbf{L}^t)} - \underbrace{\sum_{(i,j)\in\mathcal{N}_2} \log p\left(l_i^t, l_j^t \mid \vec{z}_i^t, \vec{z}_j^t\right)}_{:=\log \mathcal{B}(\mathbf{L}^t|\mathcal{Z}^t)}, \tag{3}$$

exploiting $p\left(\vec{z}_i^t, \vec{z}_j^t \mid l_i^t, l_j^t\right) \cdot p\left(l_i^t, l_j^t\right) \propto p\left(l_i^t, l_j^t \mid \vec{z}_i^t, \vec{z}_j^t\right)$. Next the hidden object parameter set $\mathbf{\Theta}$ is introduced describing the state of the K moving objects in the scene, that is the position of the j-th object described by its geometric center X_j and Z_j, the object velocity V_{xj} and V_{zj}, and its object dimensions, namely the object width $|\Delta X_j|$, height H_j and length $|\Delta Z_j|$:

$$\begin{aligned} &\mathbf{\Theta} = \{\Theta_1, \ldots, \Theta_K\} \text{ and} \\ &\Theta_j = \{X_j, Z_j, V_{xj}, V_{zj}, |\Delta X_j|, H_j, |\Delta Z_j|\}. \end{aligned} \tag{4}$$

This way, the global energy function in Equation 3 becomes

$$E = -\log Q\left(\mathbf{L}^t\right) - \log p\left(\mathbf{L}^{t-1} \mid \mathbf{L}^t\right) - \log \int_{\mathbf{\Theta}} Q\left(\mathcal{Z}^t, \mathbf{\Theta} \mid \mathbf{L}^t\right) \mathrm{d}\mathbf{\Theta} - \log \mathcal{B}\left(\mathbf{L}^t \mid \mathcal{Z}^t\right). \tag{5}$$

Applying Bayes' theorem and Taylor expanding the integrand using the Laplace method [26] yields

$$\begin{aligned} \log Q\left(\mathcal{Z}^t, \mathbf{\Theta} \mid \mathbf{L}^t\right) &= \log Q\left(\mathcal{Z}^t \mid \mathbf{L}^t, \mathbf{\Theta}\right) Q\left(\mathbf{\Theta} \mid \mathbf{L}^t\right) \\ &\approx \log Q\left(\mathcal{Z}^t \mid \mathbf{L}^t, \mathbf{\Theta}_{map}\right) Q\left(\mathbf{\Theta}_{map} \mid \mathbf{L}^t\right) \\ &\quad - \frac{1}{2}\left(\mathbf{\Theta} - \mathbf{\Theta}_{map}\right)^T \mathbf{A} \left(\mathbf{\Theta} - \mathbf{\Theta}_{map}\right) + \ldots, \end{aligned} \tag{6}$$

where $\mathbf{\Theta}_{map}$ denotes the value of $\mathbf{\Theta}$ at the mode of the integrand and $\mathbf{A}$ is the Hessian matrix of second derivatives

$$\mathbf{A} = -\nabla\nabla \log Q\left(\mathcal{Z}^t \mid \mathbf{L}^t, \mathbf{\Theta}_{map}\right) Q\left(\mathbf{\Theta}_{map} \mid \mathbf{L}^t\right). \tag{7}$$

In this case, the integral given in equation 5 can be solved analytically resulting in

$$\begin{aligned} E = &-\log Q\left(\mathbf{L}^t\right) - \log p\left(\mathbf{L}^{t-1} \mid \mathbf{L}^t\right) \\ &- \log Q\left(\mathcal{Z}^t, \mid \mathbf{\Theta}_{map}, \mathbf{L}^t\right) - \log Q\left(\mathbf{\Theta}_{map} \mid \mathbf{L}^t\right) \\ &- \frac{M}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(|\mathbf{A}|\right) - \log \mathcal{B}\left(\mathbf{L}^t \mid \mathcal{Z}^t\right), \end{aligned} \tag{8}$$

where M is the parameter dimension of $\mathbf{\Theta}$ ($M = 7 \cdot K$). The determinant $|\mathbf{A}|$ can be approximated very roughly [4]

$$\frac{1}{2}\log\left(|\mathbf{A}|\right) \approx \frac{M}{2}\log\left(N\right), \tag{9}$$

assuming the prior $Q\left(\mathbf{\Theta} \mid \mathbf{L}^t\right)$ is broad and that the measurements $\mathcal{Z}^t$ originate from statistically independent degradations, i.e. the measurements are conditionally independent, given the object parameters $\mathbf{\Theta}$ [20]. In summary, the energy function

$$\begin{aligned} E = &-\underbrace{\log Q\left(\mathbf{L}^t\right)}_{\text{prior term}} - \underbrace{\log p\left(\mathbf{L}^{t-1} \mid \mathbf{L}^t\right)}_{\text{temporal consistency}} \\ &- \underbrace{\log Q\left(\mathcal{Z}^t \mid \mathbf{\Theta}_{map}, \mathbf{L}^t\right)}_{\text{data term}} \\ &- \underbrace{\frac{M}{2}\left(\log\frac{2\pi}{N}\right)}_{\text{BIC}} - \underbrace{\log \mathcal{B}\left(\mathbf{L}^t \mid \mathcal{Z}^t\right)}_{\text{smoothness term}} \end{aligned} \tag{10}$$

is minimized.

For the temporal consistency term, for almost all stixels at time step t a predecessor stixel in the previous frame $t-1$ is determined using optical flow correspondences. This term respresents a temporal class transition matrix that is estimated on the basis of ground truth material [11].

The data term given by Equation 10 is decomposed into a height term, a position term and a motion term

$$\begin{aligned} p\left(\vec{z}_i^t \mid \mathbf{\Theta}_{map}, l_i^t\right) \approx &\underbrace{p\left(\dot{X}_i^t, \dot{Z}_i^t \mid \mathbf{\Theta}_{map}, Z_i^t, l_i^t\right)}_{\text{motion term}} \cdot \\ &\underbrace{p\left(X_i^t, Z_i^t \mid \mathbf{\Theta}_{map}, l_i^t\right)}_{\text{position term}} \cdot \\ &\underbrace{p\left(H_i^t \mid \mathbf{\Theta}_{map}, l_i^t\right)}_{\text{height term}}. \end{aligned} \tag{11}$$

These terms were also learned from a large ground truth database, c.f. [11].

The prior term $Q\left(\mathbf{L}^t\right)$ favors the static background class. Typically, about 85% of all stixels in usual traffic scenes are stationary background [11].

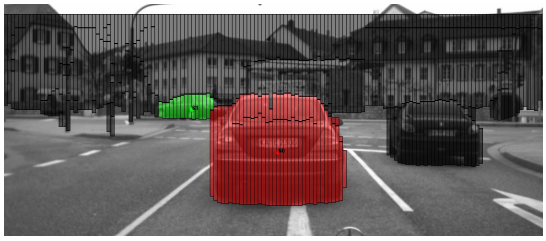
Finally, the binary term is modeled as a distance-sensitive Potts model [11].

It is computationally infeasible to optimize Equation 10 directly because $\mathbf{\Theta}$ and $\mathbf{L}^t$ are dependent on each other. For that reason, a two-stage optimization technique is used to find the maximum likelihood solution. Note that this is just a local optimum in general. A good initialization is a prerequisite to achieve good results.

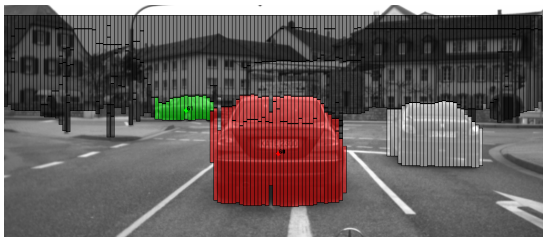
This contribution proposes two approaches to this goal: the first approach tries to estimate the hidden object parameters in an iterative manner over time, the second approach uses a radar sensor to be a source for this parameter vector. The following subsections III-A and III-B describe both approaches in more detail.

A. Vision-based Iterative Parameter Optimization

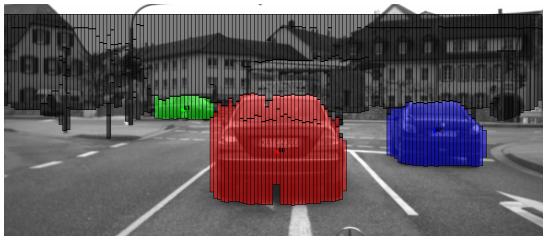
The vision-based segmentation approach alternates in an EM-like manner between segmentation cycles to find the most probable segmentation $\mathbf{L}^t$ for fixed object parameters $\mathbf{\Theta}^{t-1}$ and it re-estimates the object parameters $\mathbf{\Theta}^t$ for a fixed segmentation $\mathbf{L}^t$. Instead of iterating until convergence for a single image, the optimization is performed over several images. This way, the approach exploits the strong



(a) First image. Two moving objects are already detected, the car on the right side is starting to drive.



(b) Second image. The moving object on the right side is detected as an unknown moving object shown in white. Subsequently, its object parameter vector will be estimated as described in the main text.



(c) Third image. The unknown moving object is a known object now with its parameter vector Θ_3.

Fig. 2. Visualization of the optimization algorithm on the basis of three successive images.

correlations between neighboring images of a whole image sequence and it is considerably faster. As a result, the approach can formulate temporal expectations and it combines the segmentation step with an object tracking. See Algorithm 1 for a detailed description and Figure 2 for visualization.

Data: Dynamic Stixel World at time t
Result: Stixel Object class segmentation $\mathbf{L}^t$ and object parameter estimation $\mathbf{\Theta}^t$.
1 Compute MAP solution $\mathbf{L}^t$ using $\alpha-$ expansion graph cut for fixed $\mathbf{\Theta}^{t-1}$;
2 Extract single unknown moving objects via clustering analysis;
3 Re-estimate object parameters $\mathbf{\Theta}^t$ by gradient-descent in equation 10;

Algorithm 1: Alternating labeling and parameter estimation strategy.

In order to initialize the parameters of the moving objects, a distinction is made between *known objects*, which have been observed before and that have an already existing parameter vector Θ_j^{t-1}, and *unknown moving objects*. Unknown moving objects have not been observed so far and will change their status to a *known object* in the next frame $t+1$. The *unknown moving object* class helps to initialize a new *known object*. When a new *unknown moving object* is detected as shown in Figure 2(b), its object parameters Θ_j^t are estimated, see Equation **3** in Algorithm 1 via gradient-descent in equation 10. In the next frame $t+1$ shown in Figure 2(c), the unknown moving object is a *known object* since Θ_j^t is known.
This kind of object initialization turns out to be far more efficient and less error-prone than the usual random object parameter initialization [8].
The decision between stationary background and unknown moving objects has been described in detail in [11]. The optimal solution of this sub-problem that minimizes Equation 10 for a fixed parameter vector $\mathbf{\Theta}^t$ can be found using the multi-class alpha-expansion graph-cut scheme [5], [8].
In the case of multiple, undetected moving objects, the segmented Stixels with class *unknown moving object* might contain more than one physical objects. This information is not contained in the *unknown moving object* class. For that reason, initially a simple spatial clustering is performed once for the Stixels labeled as *unknown moving object*, in order to extract the different unknown moving objects in the scene.

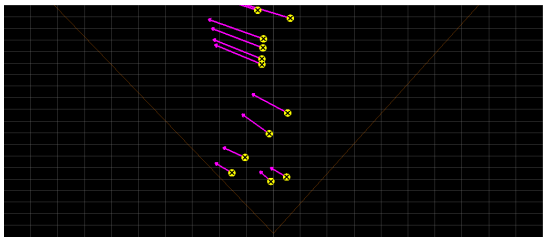
B. Radar-based Parameter Optimization

For the Radar-based object initialization strategy, the initial object state $\mathbf{\Theta}_{map}$ is obtained by an additional radar sensor instead of the detour via the unknown moving object class. After that initialization, also Algorithm 1 is executed. A radar sensor is very well suited for detecting parallel traffic, because it can directly measure such movement via Doppler Shift. However, the lateral resolution is limited in comparison to a camera system and hence the accuracy with which crossing traffic can be observed. This higher measurement uncertainty has to be taken into account in the sensor model $Q\left(\mathcal{Z}^t \mid \mathbf{\Theta}_{map}, \mathbf{L}^t\right)$. Thus it is beneficial to combine both sensors.
The radar sensor used here (Continental ARS 300) provides a large amount of object hypotheses, c.f. Figure III-B. The energy proposed in equation 10 takes into account the Bayesian Information Criterion (BIC) which penalizes the complexity of the model where complexity refers to the number M of parameters in the model. This number is proportional to the number of objects K in the scene. Taking this term into account, it is possible to find the true number of objects in the scene based on the BIC measure. The BIC decides which object hypotheses to choose, because those hypotheses reduce Equation 10 significantly without overfitting. Sensor fusion is considered in this approach as the joint optimization at sensor observation level. It is possible to take into account individual measurement uncertainties at this level.

IV. EXPERIMENTS

The experiments use a stereo camera system mounted behind the windshield of the experimental vehicle. The

(a) Radar object hypotheses. Each radar object hypothesis marked by a flipped, yellow T symbol can define the full parameter vector defined in equation 4. The BIC permits inferring the actual number of objects in the scene, rejecting the phantom hypotheses.



(b) Radar object hypotheses shown in brids-eye view relative to the ego-vehicle. The grid size is 5 m, the magenta arrows show the predicted position within the next half second.

height of the camera is 1.17 m with a base line of about 22 cm and the image resolution is 1024x440 px. The camera system records image sequences at 25 Hz. The optical flow correspondences for the stixel tracking are obtained from the well-known Kanade-Lucas-Tomasi (KLT) [24] tracker. In order to determine the required ego motion estimation, speed and yaw rate are extracted from the inertial sensors of the experimental vehicle.

The segmentation step takes about 1 ms on a single CPU core. For each frame, only one segmentation cycle consisting of a segmentation step to find $\mathbf{L}^t$ and a parameter estimation step for $\mathbf{\Theta}^t$ is performed. In order to evaluate the performance of this approach, the segmentation results were compared with a manually labeled ground truth data set. This data set contains about 80 000 images, the complete data from a test drive with a length of about one hour. The data set roughly consists half of rural roads and half of urban scenarios. All images shown in this contribution were obtained from this test drive. Every 80th image has been manually labeled to provide ground truth material as a representative sample and to avoid strong correlations between neighboring frames. In this ground truth database, there are several (stixel-wise) labeled moving objects in addition to labeled stationary background. Objects are included in the evaluation up to the detection limit of the Stixel World (about 130 m).

The experimental results are summarized in Figure 3 and Figure 4. There, the x-axis specifies the required minimum overlap: objects are considered to be segmented correctly if they overlap more than $x\%$ with a labeled object. Besides that, the figures differentiate between various distances. Figure 3 shows the detection rate of different moving objects for the vision-only based solution and in Figure 4 for the radar assisted approach. Adding the radar information

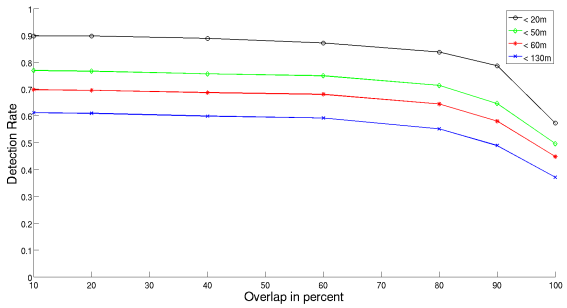


Fig. 3. Moving object detection rate based on the vision-only solution. The x-axis specifies the minimum required overlap of the segmentation result with a ground truth object. A distinction is made between different distance ranges.

increases the detection rate by about ten to fifteen percent in comparison to the vision-only solution. Especially for large distances, it is extremely difficult to separate oncoming cars from stationary background, based on their motion.

For a better grading of the results and to discuss some of the remaining error cases, see Figure 5(a) and Figure 5(b). In Figure 5(b), a pedestrian walking slowly in front of a wall is not detected but such slowly moving pedestrians appear in the ground truth. Usually, the measurement motion noise is higher than the pedestrian movement, so the pedestrian cannot be reliably detected. In future work, the intention is to take into account a pedestrian classification step in order to increase the sensitivity of the system.

If requesting for a very high overlap ($\geq 90\%$), the detection rate drops significantly. This decrease is comprehensible and corresponds to - depending on the distance - one or two stixels at the border of objects due to fluctuations in the stixel segmentation.

Complementary to this investigation, the correctly labeled stationary background (false alarms) is summarized in Table I. The low phantom rate observed in the experiments is a

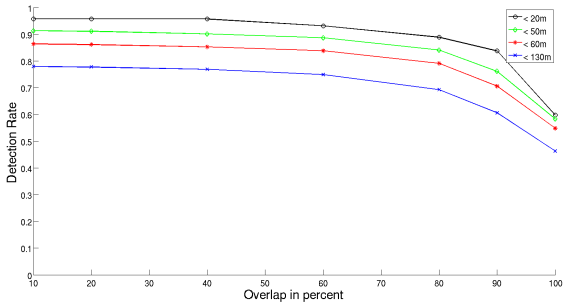
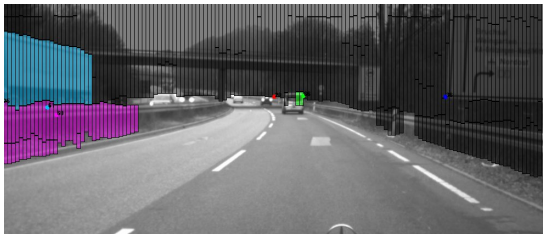


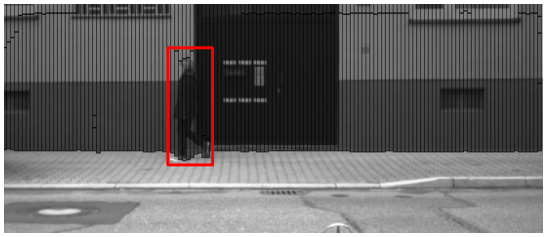
Fig. 4. Moving object detection rate based on the radar assisted solution. The x-axis specifies the minimum required overlap of the segmentation result with a ground truth object. A distinction is made between different distance ranges.

direct consequence of the strong regularization applied in this approach. See Figure 5(a) for an example of a remaining false positive detection. There are about 40 phantom stixels

for the vision-only solution, this corresponds to a phantom rate of one phantom stixel every twentieth image. Using the radar assisted approach, the phantom rate is higher, it is about one phantom stixel every six images. There are several phantom measurements, especially due to erroneous radar reflections at guardrails, c.f. Figure 5(a).



(a) Phantom example. A guardrail is incorrectly segmented as a moving object due to erroneous radar reflections and weak texture that complicates the vision-based tracking.



(b) False negative example. A slowly moving pedestrian is not separated from the stationary background.

Fig. 5. Error cases to visualize the discussion in the main text.

approach	correct background
with radar	99.18 %
without radar	99.64 %

TABLE I

THE CORRECT LABELED STATIONARY BACKGROUND STIXEL PERCENTAGE DEFINING A PHANTOM RATE.

V. CONCLUSIONS AND OUTLOOK

An EM-like CRF model for traffic scene segmentation has been presented. The difficulty of an (theoretically) uncountable infinite number of object states and classes is solved in a time-recursive fashion. The effectiveness of the proposed method has been demonstrated on the basis of ground truth data in various, challenging traffic scenes. The presented real-time capable approach has been extensively tested in the experimental vehicle.

There are further ways to develop this approach towards an increasingly powerful vision system. One intention is to take into account appearance cues, e.g. pedestrian classification. This step will help to further increase the sensitivity of the system especially for slowly moving pedestrians. Besides that, incorporating further scenario-specific knowledge from externally provided maps has the potential to yield significant improvements. Thirdly it might be beneficial to introduce a feedback loop from the object segmentation back to the stixel tracking.

REFERENCES

[1] A. Bachmann. Applying recursive em to scene segmentation. *DAGM*, 2009.

[2] A. Barth, J. Siegemund, A. Meissner, U. Franke, and W. Foerstner. Probabilistic multi-class scene flow segmentation for traffic scenes. *DAGM*, 2010.

[3] R. Benenson, R. Timofte, and L. Van Gool. Stixels estimation without depth map computation. *ICCV*, 2011.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., New York, NY, USA, 2007.

[5] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *ICCV*, 1999.

[6] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. *ECCV*, 2008.

[7] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Combining appearance and structure from motion features for road scene understanding. *BMVC*, 2009.

[8] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *IJCV*, 2012.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, B39(1), 1-38*, 1977.

[10] M. Enzweiler, M. Hummel, D. Pfeiffer, and U. Franke. Efficient stixel-based object recognition. *IV*, 2012.

[11] F. Erbs, B. Schwarz, and U. Franke. Stixmentation - probabilistic stixel based traffic scene labeling. *BMVC*, 2012.

[12] A. Ess, T. Mueller, H. Grabner, and L. van Gool. Segmentation-based urban traffic scene understanding. *BMVC*, 2009.

[13] W. Feng, J. Jia, and Z.-Q. Liu. Self-validated labeling of markov random fields for image segmentation. *TPAMI*, 2010.

[14] U. Franke, C. Rabe, H. Badino, and S. Gehrig. 6d vision - fusion of motion and stereo for robust environment perception. *DAGM Symposium*, 2005.

[15] S. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. *ICCV*, 2009.

[16] H. Hirschmueller. Accurate and efficient stereo processing by semiglobal matching and mutual information. *CVPR*, 2005.

[17] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *TPAMI*, 2004.

[18] L. Ladicky, P. Sturgess, K. Alahari, C. Russel, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. *ECCV*, 2010.

[19] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

[20] R. Mester. A bayesian view on matching and motion estimation. *SSIAI*, 2012.

[21] D. Pfeiffer. *The Stixel World: A Compact Medium-level Representation for Efficiently Modeling Dynamic Three-dimensional Environments*. PhD thesis, Humboldt-University of Berlin, Berlin, Germany, 2012.

[22] D. Pfeiffer and U. Franke. Efficient representation of traffic scenes by means of dynamic stixels. *IV*, 2010.

[23] D. Pfeiffer and U. Franke. Towards a global optimal multi-layer stixel representation of dense 3d data. *BMVC*, 2011.

[24] J. Shi and C. Tomasi. Good features to track. *CVPR*, 1994.

[25] D. Singaraju, L. Grady, , and R. Vidal. P-brush: Continuous valued mrfs with normed pairwise distributions for image segmentation. *CVPR*, 2009.

[26] D. S. Sivia. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, USA, 1996.

[27] D. Sun, E. Sudderth, and M. J. Black. Layered segmentation and optical flow estimation over time. *CVPR*, pages 1768–1775, 2012.

[28] M. F. Tappen. Utilizing variational optimization to learn markov random fields. *CVPR*, 2007.

[29] M. Unger, M. Werlberger, T. Pock, and H. Bischof. Joint motion estimation and segmentation of complex scenes with label costs and occlusion modeling. *CVPR*, 2012.

[30] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. *CVPR*, 2007.