Stixmentation - Probabilistic Stixel based Traffic Scene Labeling

Friedrich Erbs
friedrich.erbs@daimler.com
Beate Schwarz
beate.schwarz@daimler.com
Uwe Franke
uwe.franke@daimler.com

Image Understanding
Daimler AG
Boeblingen, Germany

Abstract

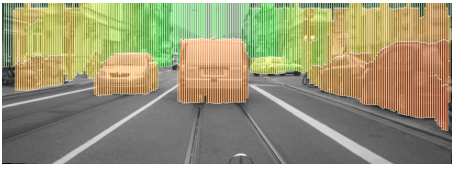
The detection of moving objects like vehicles, pedestrians or bicycles from a mobile platform is one of the most challenging and most important tasks for driver assistance and safety systems. For this purpose, we present a multi-class traffic scene segmentation approach based on the Dynamic Stixel World, an efficient super-pixel object representation. In this approach, each Stixel is assigned either to a quantized maneuver motion class like oncoming, or left-moving or to static background. The formulation integrates multiple 3D and motion features as well as spatio-temporal prior knowledge in a probabilistic conditional random field (CRF) framework. The real-time capable method is evaluated quantitatively in various challenging, cluttered urban traffic scenes. The experimental results yield highly accurate segmentation of urban traffic scenarios without the need for any manual parameter adjustments.

1 Introduction

Traffic scene understanding and image segmentation are active fields of research in computer vision. More and more, the concepts developed in these fields culminate in impressive autonomous driving applications such as the Google self driving-car [29], AnnieWAY [23] or Stadtpilot [14]. For these applications, detecting and tracking other traffic participants such as vehicles, bicyclists or pedestrians is of particular interest. For this purpose, the above mentioned projects use a high-precision laser scanner on top of the experimental vehicle. Our goal is to realize such autonomous driving using a low cost stereo camera system.
To this end, we use the *Dynamic Stixel World* [25, 26] as input to our approach, as shown in Figure 1. The Stixel World is an efficient super-pixel representation of 3D traffic scenes. The relevant information in the scene is represented with a few hundreds Stixels instead of thousands of individual stereo depth measurements. This compression of the input data volume also reduces the computational burden for our segmentation step by at least three orders of magnitude, thus enabling real-time capability. Besides that, the Stixel World is insensitive to outliers which boosts the robustness of subsequent algorithms.
This work presents a probabilistic conditional random field framework for segmenting moving objects into different motion classes. The main steps of our segmentation process are

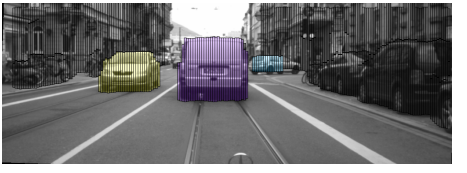
© 2012. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.
http://dx.doi.org/10.5244/C.26.71

(a) SGM Stereo [10, 13]. The color represents the distance to the obstacle with red being close and green far away.



(b) The multi-layered Stixel World [26] yields the drivable freespace information and approximates the object boundaries. The Stixel width was set to $w = 5$ px.

(c) Dynamic Stixel World [25]. The arrows point to the predicted Stixel position within the next half second.



(d) Segmentation result with three moving objects shown in yellow, magenta and cyan. The static background is shown in black.

Figure 1: Example results for the different steps of our segmentation process chain.

summarized in Figure 1. It starts from dense stereo depth maps obtained by the Semi-Global Matching (SGM) stereo algorithm [10, 13] as shown in Figure 1(a). Then, the multi-layered Stixel World [26] (Figure 1(b)) and the Dynamic Stixel World [25] (Figure 1(c)) are computed. The final segmentation result depicted in Figure 1(d) separates the image into different motion classes. These include oncoming, forward-moving, right-moving and static background (shown in yellow, magenta, cyan and black respectively). A short video example provided as supplementary material gives a more detailed indication of the approach and further illustrates the different steps of the segmentation pipeline.

The remainder of this paper is structured as follows: Section 2 briefly points out related work. Section 3 introduces the mathematical modeling of our segmentation framework. Section 4 discusses the input feature probability distributions that we have built up from a ground truth database from a 25-minute urban drive. These probability distributions are the most important component of the segmentation process. Furthermore, Section 4 introduces the spatio-temporal couplings between neighboring Stixels. The performance of the segmentation approach has been evaluated on a different 5-minute drive through another urban environment. Parts of the corresponding results are presented in Section 5. Finally, Section 6 concludes this contribution.

2 Related Work

Image segmentation has been tackled successfully as a multi-class labeling problem mainly for static environments, e.g. by [8, 31] using color or texture cues.

However, such appearance-based features suffer from difficulties in the presence of strongly varying lighting or weather conditions. Therefore, several approaches use structure from motion derived 3D world information to classify static scenes, e.g. [4, 5].

Liu *et al*. [21] propose a label transfer framework using SIFT flow correspondences between

a testing image and its best matching image in a large, labeled database. In [17], Ladicky *et al*. consider dense stereo reconstruction and object segmentation for static scenes as well in a joint optimization framework and use height information to couple both approaches.
Recently, there is a trend to use super-pixels for image segmentation, e.g. in [16]. In [32, 33], the authors propose to use super-pixels for segmentation in order to obtain more discriminating features defined over larger image regions and to reduce the computational costs.
A key criterion for super-pixels is their consistency with the actual object boundaries which can turn out to be difficult, especially for cluttered images. In this context, stereo depth information has proven to be a valuable feature for an accurate super-pixel extraction, such as in case of the Stixel World [26]. Furthermore, Stixels have also been successfully used as an attention guide for object classifiers in [2, 7].
Besides using stereo depth information, recent progress in scene flow computation allows considering motion information as a powerful cue to discriminate between different objects which cannot be separated based on depth or appearance information alone [22, 27].
Thus, this contribution relies solely on such depth and motion information to perform scene segmentation for a restricted set of object classes. The authors of [30] use graph cut to identify moving objects based on a probabilistic threshold. In [1], Barth *et al*. present a probabilistic multi-class traffic scene segmentation approach purely based on dense depth and motion information.
This work focuses on the latter approach, which it expands in various aspects. Firstly, it proposes using the Dynamic Stixel World instead of dense stereo and scene flow data. This significantly increases stability and reduces the computational burden by roughly three orders of magnitude. Secondly, it replaces the direct modeling of unary potential terms by a Bayesian approach, i.e. it actually learns class histograms from a large dataset containing 38,000 images and about ten million Stixels. Thirdly, it introduces a temporal coupling to extend the proposed single frame segmentation and to enforce temporal consistent segmentation results.

3 General Segmentation Framework

Given a stereo image sequence $\mathcal{I}$, the multi-layered Stixel World as proposed in [24] is computed first. This step partitions the input image $I^t \in \mathcal{I}$ at time step t column-wise into several layers of one of the three classes $\mathcal{C}_{\text{Stixel}} \in \{\text{street, obstacle, sky}\}$. The focus is on obstacle Stixels and the other classes street and sky are left unchanged.
However, the sole Stixel data is not yet sufficient to describe dynamic objects, which is the main objective. Therefore, in the next step, the Stixels are tracked over time by fusing stereo and optical flow information. This task follows the 6D-Vision principle proposed by Franke *et al*. in [9, 25]. In order to be able to derive absolute velocities, the motion of the ego-vehicle, measured by the inertial sensors of the experimental vehicle, is compensated.
To sum up, each Stixel with index i is defined by 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, and its velocity $\{\dot{X}_i^t, \dot{Z}_i^t\}$. Moving objects such as cars or bicycles are assumed to move on the ground plane, so it is sufficient to estimate a 2D motion vector. These five observations form a feature vector for each Stixel, $\vec{z}_i^t = \{\dot{X}_i^t, \dot{Z}_i^t, X_i^t, H_i^t, Z_i^t\}^{\mathrm{T}}$.
Now, let $L^t = \{l_1^t, \ldots, l_N^t\}^{\mathrm{T}}$ denote a labeling for a given input image I^t containing N dynamic Stixels. Since the focus is on separating moving and stationary objects, the attempt is made to assign each Stixel with index i to a particular class $l_i^t \in \{\text{moving, stationary}\}$. The moving

object class is further quantized into four driving behaviors: forward-moving, oncoming, left-moving and right-moving. This definition allows to separate objects that are located close together but have a different moving direction. At a first glance, the class choice might seem rather unspecific. However, this decision turns out to be the most relevant for the application and a more detailed scene description is often not required.
The fast alpha-expansion multi-class graph cut optimization scheme described in [3] is used as the inference step.
The feature vectors of all N Stixels in the image I^t are combined in a measurement array $\mathcal{Z}^t = \{\vec{z}_1^t, \ldots, \vec{z}_N^t\}$. Given this measurement array $\mathcal{Z}^t$ and given the labeling result from the previous time step L^{t-1}, allows inferring the most probable labeling defined as $\arg\max_{L \in \mathcal{L}} p\left(L^t \mid \mathcal{Z}^t, L^{t-1}\right)$ from the set of all possible labelings $\mathcal{L}$. This probability is modeled as a conditional random field [13] with a maximum clique size of two. The most probable labeling minimizes the following log-likelihood energy E [12]

$$\begin{aligned} E &= -\log p\left(L^t \mid \mathcal{Z}^t, L^{t-1}\right) \\ &\sim \sum_{i=1}^{N} \psi\left(l_i^t \mid \mathcal{Z}^t, L^{t-1}\right) + \lambda \cdot \sum_{(i,j)\in\mathcal{N}_2} \phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right). \end{aligned} \tag{1}$$

In this context, $\mathcal{N}_2$ denotes the set of all neighboring Stixels and the term λ is a scaling factor for the binary term $\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right)$. The unary terms are modeled

$$\begin{aligned} \psi\left(l_i^t \mid \mathcal{Z}^t, L^{t-1}\right) &= -\log p\left(l_i^t \mid \mathcal{Z}^t, l_i^{t-1}\right), \text{ where} \\ p\left(l_i^t \mid \mathcal{Z}^t, l_i^{t-1}\right) &= p\left(l_i^t \mid \vec{z}_i^t, l_i^{t-1}\right) \\ &\propto p\left(\vec{z}_i^t, l_i^{t-1} \mid l_i^t\right) \cdot p\left(l_i^t\right) \\ &\approx \underbrace{p\left(\vec{z}_i^t \mid l_i^t\right)}_{\text{Data Term}} \cdot \underbrace{p\left(l_i^{t-1} \mid l_i^t\right)}_{\text{Temporal Expectation}} \cdot \underbrace{p\left(l_i^t\right)}_{\text{Prior Term}}. \end{aligned} \tag{2}$$

The unary potential terms are defined to be the negative log-likelihoods of the statistical probabilities. The following section discusses in detail the statistical modeling of the unary terms $\psi\left(l_i^t \mid \mathcal{Z}^t, L^{t-1}\right)$ and the binary regularization term $\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right)$.

4 Feature Selection

Clearly, the velocity measurements are the most important feature to separate moving objects from stationary background. The velocity distributions for moving objects and for static background in typical urban traffic scenes are set up from the training ground truth dataset containing manually labeled Stixels as training examples. Modeling

$$\begin{aligned} p\left(\vec{z}_i^t \mid l_i^t\right) &= p\left(\dot{X}_i^t, \dot{Z}_i^t, X_i^t, H_i^t, Z_i^t \mid l_i^t\right) \\ &\approx \underbrace{p\left(\dot{X}_i^t, \dot{Z}_i^t \mid Z_i^t, l_i^t\right)}_{\text{motion term}} \cdot \underbrace{p\left(X_i^t, Z_i^t \mid l_i^t\right)}_{\text{position term}} \cdot \underbrace{p\left(H_i^t \mid l_i^t\right)}_{\text{height term}}. \end{aligned} \tag{3}$$

In the first motion term, the dependence on Z_i^t is important because for a stereo camera sensor the distance uncertainty grows quadratically. The farther away a Stixel is, the larger is its motion uncertainty. Distance dimension Z is quantized to keep the learning step feasible, see

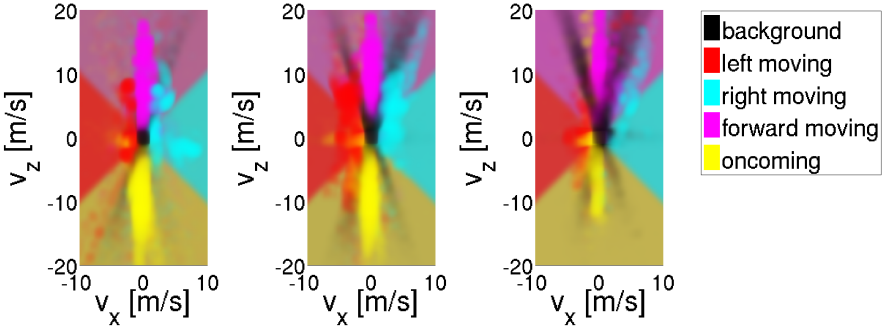
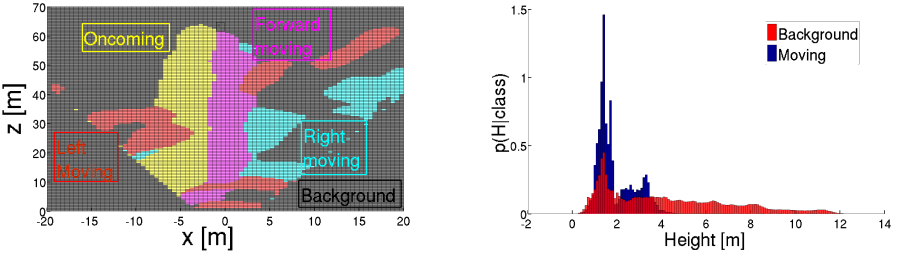


Figure 2: The probability $p\left(\dot{X}_i^t, \dot{Z}_i^t \mid l_i^t, Z_i^t\right)$ is color encoded for different distance ranges Z_i^t. On the left side, $Z_i^t \in \{0-20\}$ m, in the middle, $Z_i^t \in \{20-40\}$ m, and on the right side $Z_i^t \in \{40-70\}$ m is shown.

Figure 2 for an illustration. As shown in this graph, the background motion distribution is spread more for larger distances. Hence, it becomes quite difficult to separate slow moving objects from stationary background. Usually, the background distribution is modeled to be Gaussian and its variance is estimated using error propagation from estimated scene flow confidences, cf. [20, 30]. However, this assumption does not hold for the present setup, as can be seen from Figure 2, and thus yields inferior results.

Figure 3(a) visualizes the probability of occurrence of the object classes at different world positions $\{X_i^t, Z_i^t\}$. Note that the most probable object class l_i^t at different world positions are



(a) The most probable class l_i for different Stixel world positions $\{X_i, Z_i\}$ is color encoded. The colors are explained in the figure.

(b) $p(H_i \mid l_i)$ for moving objects and the static background class. The overlapping area is marked using dark red.

Figure 3: Positional (3(a)) and height (3(b)) probability distributions for the five object classes.

color-coded, i.e. $\arg\max_{l_i^t} p(l_i^t \mid X_i^t, Z_i^t)$, instead of the likelihood $p(X_i^t, Z_i^t \mid l_i^t)$ itself. This helps to keep the visualization uncluttered. The ego vehicle is placed at the origin of the underlying (X, Z) coordinate system.

The underlying positional distributions reflect various traffic related aspects. Typically, oncoming cars are located to the left of the ego vehicle. Stixels in front are often forward-moving due to a leading vehicle. Furthermore, Stixels close to the image border are often stationary background.

Finally, the height term favors very high Stixels to be stationary background. This is just natural, because those often model buildings or other tall infrastructure. Stixels modeling cars, bicycles and pedestrians have rather moderate heights. We have evaluated the underlying height statistics as shown in Figure 3(b).

For the temporal expectation term, for each Stixel at time step t a predecessor Stixel is determined using optical flow analysis. The temporal object class consistency $p\left(l_i^{t-1} \mid l_i^t\right)$ was evaluated in our training data set. Given the label l_i^t for each Stixel, the resulting class label from the previous time step l_i^{t-1} was analyzed. In most cases, we have decided for the correct ground truth label. This consideration defines the transition matrix $p\left(l_i^{t-1} \mid l_i^t\right)$. The statistical findings are summarized in Table 1. The temporal expectation term favors a consistent label decision. Nevertheless, it is worth pointing out that it might also cause unwanted low-pass effects.

GT / predecessor class	BG	LEFT	RIGHT	FW	ON
BG	95.57	8.39	16.73	7.91	9.16
LEFT	0.06	73.72	1.69	0.87	1.65
RIGHT	0.07	2.16	70.23	1.47	0.04
FW	3.60	1.74	10.07	89.56	0.15
ON	0.70	13.98	1.28	0.19	89.00

Table 1: The old class decision, l_i^{t-1} is considered to be a prior for the current segmentation l_i^t. This figure shows the statistical transition probabilities $p\left(l_i^{t-1} \mid l_i^t\right)$ in percent.
BG = background, LEFT = left-moving object, RIGHT = right-moving object
FW = forward-moving object and ON = oncoming object.

According to the training data, the prior term $p\left(l_i^t\right)$ favors the static background class. In typical urban traffic scenes, roughly 87% of all Stixels are stationary. The remaining prior class probabilities are shown in the first line of Table 2.

The smoothness term $\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right)$ is modeled as a Potts model, this way favoring neighboring Stixels to belong to the same class. Each Stixel is modeled to be a node in the CRF. The maximum clique size is restricted to two, and thus only nearest neighbor Stixel interactions are considered. Since there is no regular underlying grid, each Stixel can have an arbitrary number of neighboring Stixels.

The spatial correlation between neighboring Stixels has been investigated using the training data set. Evidently, it strongly depends on their relative depth difference. The closer two Stixels are, the more likely they belong to the same object. Since the focus is on working with stereo data, disparity deviations are considered directly rather than depth differences. In this context, the underlying disparity uncertainty σ_d is modeled to be constant [28]. Additionally, it has to taken into account that objects (e.g. engine hoods) often do not perfectly fulfill the constant depth assumption, which is inherently assumed by the Stixel model. It is therefore necessary to consider both the disparity uncertainty σ_d and this model violation σ_{world} in a joint metric $\Delta_{\text{disp}} = \frac{\|d_i - d_j\|}{\sigma_{\text{disp}}}$ between two neighboring Stixels i and j. Hereby, σ_{disp} is defined as $\sigma_{\text{disp}} = \max\left(\sigma_d, \sigma_{\text{world}}\right)$. Accordingly, the smoothness term $\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right)$ is modeled as

$$\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right) = \begin{cases} -\log\left(\mathrm{p}_{\text{equal}}\left(\Delta_{\text{disp}}\right)\right), & \text{if } l_i^t = l_j^t \\ -\log\left(1 - \mathrm{p}_{\text{equal}}\left(\Delta_{\text{disp}}\right)\right), & \text{else.} \end{cases} \tag{4}$$

The evaluation results of the class correlation between neighboring Stixels in the training

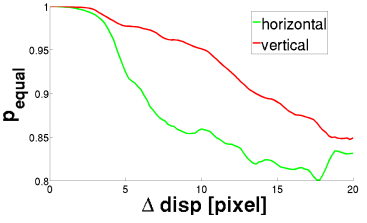


Figure 4: The spatial correlation between vertical (red) and horizontal (green) neighboring Stixels is plotted as a function of their mutual distances as elaborated in the text. Note that this figure shows a class correlation between neighboring Stixels, the neighboring Stixels do not necessarily need to belong to the same physical object.

data set are illustrated in Figure 4. In this Figure, the vertical and horizontal neighboring Stixels are separated. It is shown that, typically, vertical adjacent Stixels are more correlated than horizontal neighbors which justifies the anisotropic modeling. The work in [15] showed that in order for the fast alpha-expansion-move algorithm to be applicable, for all labels α, β, and γ

$$\phi\left(\alpha,\alpha \mid \mathcal{Z}^t\right)+\phi\left(\beta,\gamma \mid \mathcal{Z}^t\right) \leq \phi\left(\alpha,\gamma \mid \mathcal{Z}^t\right)+\phi\left(\beta,\alpha \mid \mathcal{Z}^t\right) \tag{5}$$

must hold which is clearly valid for the modeling shown in Figure 4.
In order to find the best value for the scaling parameter λ, its value is discretized and the global segmentation error is sampled on the training data set for these values, cf. Figure 5. Taking the value which yields the smallest error as a good initial guess, the estimated best

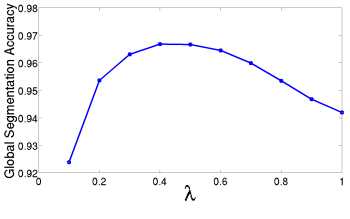


Figure 5: Influence of the scaling parameter λ on the global segmentation accuracy in our training data set. Setting $\lambda = 0.4$ yields the highest accuracy.

λ is refined using Newton-Raphson. However, as can be seen from Figure 5, the global segmentation accuracy depends only weakly on the exact value of this parameter. Figure 6 color-codes the resulting coupling strength from the class correlations shown in Figure 4 between adjacent Stixels. Red corresponds to a strong coupling and green denotes a weak coupling.

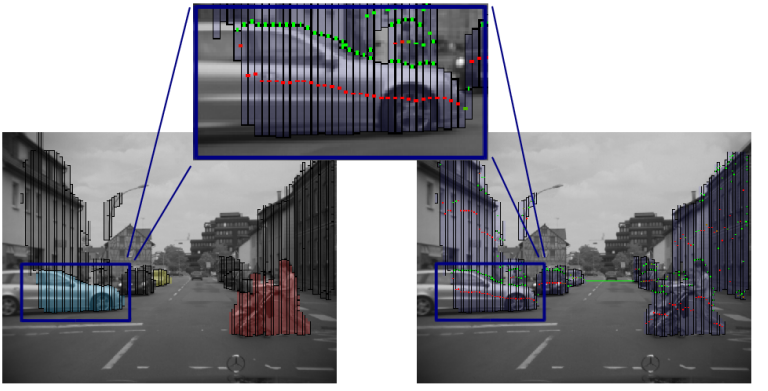


Figure 6: Spatial couplings between Stixels. The coupling strength is color encoded where red symbolizes strong coupling and green corresponds to a weak coupling.

Features	BG	LEFT	RIGHT	FW	ON	**Average**	**Global**
prior (GT)	87.56	0.73	0.73	7.49	3.49	-	-
All	99.54	85.07	95.65	79.77	83.18	88.64	98.01
w/o motion	92.26	0.00	0.00	48.04	56.11	39.28	85.40
w/o height	99.69	83.68	94.91	77.40	80.63	87.26	97.95
w/o prior	88.69	93.01	97.27	89.01	88.86	91.37	88.99
w/o position	99.83	92.70	95.80	75.73	66.42	86.10	97.98
w/o binary	98.14	80.25	89.72	75.55	80.91	84.91	96.31
w/o temporal	99.70	84.30	95.17	77.96	83.08	88.04	98.06

Table 2: Stixel-wise percentage accuracy for our evaluation sequence. BG = background, LEFT = left-moving object, RIGHT = right-moving object, FW = forward-moving object and ON = oncoming object. "Global" denotes the percentage of Stixels that were correctly classified, "Average" is the average of the per-class accuracies.

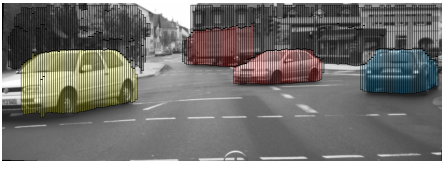
5 Experimental Results

The experiments use a stereo camera system mounted behind the windshield of the experimental vehicle. The height of the camera system is 1.17 m with a base line of 22 cm as well as an image resolution of 1024x440 px.

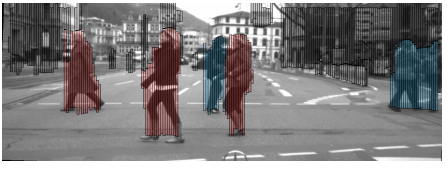
The dense stereo depth maps are computed in real-time at 25 Hz on a dedicated FPGA platform using the Semi-Global Matching algorithm as described in [10, 13]. The segmentation algorithm has been implemented in C++ and takes about 1 ms on a single CPU core.

To test the performance of the approach, the segmentation results were compared with a manually labeled data set, containing about 8000 images recorded from our experimental vehicle. To our best knowledge, at the time of this publication, there was only very little ground truth material available for traffic scene related stereo camera sequences. The only publically available database for stereo image sequences was the Leuven Moving Vehicle Sequence [6, 19] with ground truth material for about 80 images which is insufficient to learn the proposed probability distributions. For that reason we decided to setup a new own

ground truth database. In the future we are planning to also use the recently published KITTI Vision Benchmark Suite [11] which is of comparable size.
All experiments have been performed with a single parameter set, and thus without any manual parameter tuning.



(a) Example result from the training data set.



(b) Example scene of our evaluation sequence.

Figure 7: The training (left) and evaluation (right) sequence for our framework.

The performance of the system is summarized in Table 2. Distinct features have been omitted in order to test their influence on the final segmentation result.
As it turns out, the best overall performance is achieved taking into account all the proposed features from Section 4 and leaving out the temporal term. In this case, the average labeling accuracy is 98.06%. However, when taking account the temporal coherence constraint, the results are quite similar with 98.01%. The positive influence of the temporal constraint is canceled by unwanted low-pass effects.
The motion cue turns out to be the most discriminative feature. By ignoring this term, the global performance decreases to 85.40%, as shown in Table 2. In this case, some maneuvering classes, such as right-moving, are not classified at all. Still, as one can see from the results, the other features also play an important role for the global segmentation result. The prior term, for example, proves to be important because it suppresses phantom objects, which are wrong moving objects as a result of motion artifacts. By leaving this term out, the overall labeling accuracy drops to 88.99%, even though the average per-class accuracy increases because the static background class is not favored any longer.
As one might expect, the position term turns out to be advantageous especially for oncoming objects and objects driving ahead. In this case, the segmentation accuracy rises significantly for these classes.

6 Conclusions and Outlook

This contribution introduced a generic framework for detecting and segmenting moving objects based on the Dynamic Stixel World. The key conclusion from the experiments is that learning statistical relations from sufficient training data sets yields a powerful and robust segmentation apparatus with no need for any manual parameter tuning. As shown by the results, the approach generalizes unseen new traffic scenes well.
This work focused primarily on urban traffic scenes. However, this approach can easily be adapted to other scenarios such as highways or rural roads. Taking into account scenario specific knowledge proves to be highly advantageous.
Using the Stixel World instead of dense stereo and pixel-wise motion information yields significant improvements with respect to stability and real-time capability because the amount of input data is reduced considerably and the Stixel optimization step is insensitive to outliers. Errors due to a wrong Stixel segmentation were found to be very seldom. If at all,

problems arose due to wrong motion information and due to insufficient training examples. In future work, the intention is to take into account appearance cues or to consider different traffic scenarios. Besides that, incorporating further scenario-specific knowledge from externally provided maps has the potential to yield significant improvements.

(a) Highway scenario with several phantom objects on the left side. This result is obtained using the urban model.

(b) Segmentation result when using a specific highway model.

Figure 8: Example result of a highway scenario which clearly reveals the necessity of scenario-specific regularization.

Ongoing work is focusing on learning scenario-specific models.
Figure 8(a) shows the result obtained if the urban model is applied to a highway scene. Due to noisy motion measurements on the guard rail, the optimization erroneously decides for a moving object. However, if a specific highway model is used, one gets the result shown in Figure 8(b). The intention is to exploit this property of the system to reduce segmentation errors in case of adverse weather conditions.

References

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

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

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

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

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

[6] N. Cornelis, B. Leibe, K. Cornelis, and L. Van Gool. 3d urban scene modeling integrating recognition and reconstruction. *IJCV*, 2007.

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

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

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

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

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *CVPR*, 2012.

[12] Robert M. Gray. *Entropy and information theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.

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

[14] Technical University of Braunschweig Institute of Control Engineering, Institute of Flight Guidance. Stadtpilot. *http://stadtpilot.tu-bs.de/en/stadtpilot/*, 2007.

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

[16] L. Ladicky, C. Russel, P. Kohli, and P. Torr. Associative hierarchical crfs for object class image segmentation. *ICCV*, 2009.

[17] L. Ladicky, P. Sturgess, C. Russel, S. Sengupta, Y.Bastanlar, W. Clocksin, and P. H. S. Torr. Joint optimisation for object class segmentation and dense stereo reconstruction. *BMVC*, 2010.

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

[19] B. Leibe, , N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. *CVPR*, 2007.

[20] P. Lenz, J. Ziegler, A. Geiger, and M. Roser. Sparse scene flow segmentation for moving object detection in urban environments. *IV*, 2011.

[21] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2368–2382, 2011.

[22] T. Mueller, J. Rannacher, C. Rabe, and U. Franke. Feature- and depth-supported modified total variation optical flow for 3d motion field estimation in real scenes. *CVPR*, 2011.

[23] Department of Measurement and Karlsruhe Institute of Technology Control. Annieway. *http://www.mrt.kit.edu/annieway/*, 2006.

[24] 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.

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

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

[27] C. Rabe, T. Mueller, A. Wedel, and U. Franke. Dense, robust, and accurate 3d motion field estimation from stereo image sequences in real-time. *ECCV*, 2010.

[28] Clemens Rabe. *Detection of Moving Objects by Spatio-Temporal Motion Analysis*. PhD thesis, University of Kiel, Kiel, Germany, 2011.

[29] S. Thrun. What we're driving at. *http://googleblog.blogspot.de/2010/10/what-were-driving-at.html*, 2010.

[30] A. Wedel, A. Meissner, C. Rabe, U. Franke, and D. Cremers. Detection and segmentation of independently moving objects from dense scene flow. *EMMCVPR*, 2009.

[31] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. *ECCV*, 2008.

[32] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. *ICCV*, 2009.

[33] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. *ECCV*, 2010.