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

The detection and segmentation 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 for traffic scenes.

Related works often performs pixel-wise segmentation and is mainly focused on static scenes. Using the Stixel World creates dramatic advantages in comparison with such traditional approaches. The relevant information in the scene is represented with a few hundreds Stixels instead of hundreds of thousands of individual dense stereo depth and optical flow measurements. This compression of the input data volume also reduces the computational burden for a subsequent segmentation step by at least three orders of magnitude, thus enabling real-time capability. Besides that, the Stixel World turns out to be extremely stable with respect to outliers due to a global optimization used in its calculation. Subsequent algorithms profit strongly from this high reliability of data. Taking into account motion information creates the possibility to discriminate between different objects which cannot be separated based on depth or appearance information alone.

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 summarized in Figure 1. It starts from dense stereo depth maps obtained by the Semi-Global Matching (SGM) stereo algorithm [1] as shown in Figure 1(a). Then, the multi-layered Stixel World [3] and the Dynamic Stixel World which is extended by motion information [2] (Figure 1(b)) are computed. The final segmentation result depicted in Figure 1(c) 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).

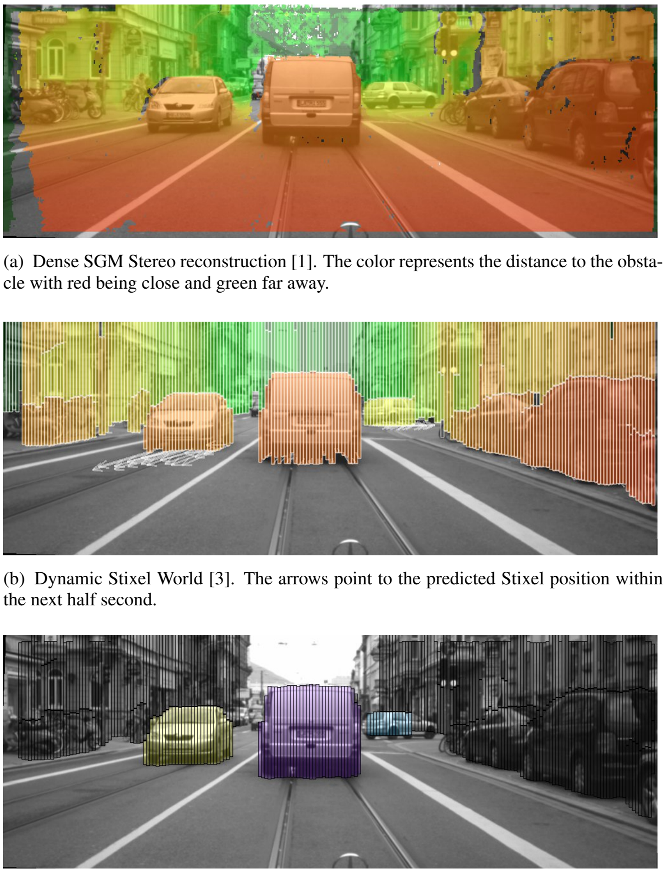
For the segmentation, we seek the most probable labeling minimizes the following log-likelihood energy E

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

In this context, $L^t = \{l_1^t, \dots, l_N^t\}^{\mathrm{T}}$ denotes a labeling for a given input image I^t containing N dynamic Stixels and the observations for all Stixels are combined in a measurement array $\mathcal{Z}^t = \{\vec{z}_1^t, \dots, \vec{z}_N^t\}$. $\mathcal{N}_2$ 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

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

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. The unary potential terms are defined to be the negative log-likelihoods of statistical probability distributions. These distributions for the different object classes in typical urban traffic scenes were set up from a large training ground truth database containing manually labeled Stixels as training examples. This database contains about 38,000 images and about ten million Stixels. All parameters of the energy function defined in 1 including the weight parameter λ were learned from this dataset.



(a) Dense SGM Stereo reconstruction [1]. The color represents the distance to the obstacle with red being close and green far away.

(b) Dynamic Stixel World [3]. The arrows point to the predicted Stixel position within the next half second.

(c) 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.

In order to evaluate the performance of the presented approach, the segmentation results were compared with another challenging data set, containing about 8000 images recorded from our experimental vehicle. All experiments have been performed with a single parameter set, and thus without any manual parameter tuning. The experimental results yield highly accurate segmentation of urban traffic scenarios, the average labeling accuracy is 98.06%. Additionally distinct features have been omitted in order to test their influence on the final segmentation result. Using the Stixel World allows to compute the alpha-expansion graph cut inference in real time in 1 ms on a single CPU core. 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.

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

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

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