Towards a Global Optimal Multi-Layer Stixel Representation of Dense 3D Data

David Pfeiffer
david.pfeiffer@daimer.com
Uwe Franke
uwe.franke@daimler.com

Image Understanding
Daimler AG
Böblingen, Germany

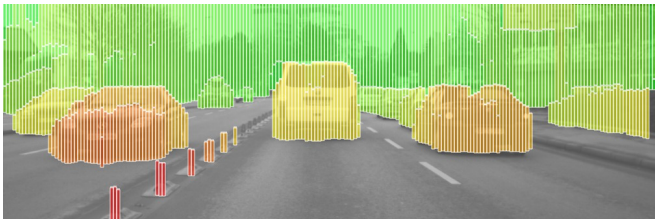

Figure 1: The multi-layer *Stixel World* result as output of the optimization. The captured scene is segmented into planar *Stixel* segments that correspond to either ground or object. The color represents the distance to the obstacle with red being close and green far away. Grey pixels belong to the ground surface.

Dense 3D data as delivered by stereo vision systems, modern laser scanners or time-of-flight cameras such as PMD is a key element for 3D scene understanding. Recent progress in stereo vision allows for energy-efficient FPGA and ASIC hardware solutions that compute high-quality dense stereo depth maps in real-time. This raises general demands for new processing schemes and medium level-representations, since applications originating from GIS, robotics, or driver assistance often can not afford to evaluate every single depth measurement individually.

In a first attempt, we addressed this task by means of the Stixel World. Today, this representation is extended, such that objects are allowed to be located at multiple depths in a column. An example representation is depicted in Figure 1.

In a unified, probabilistic approach to compute the Stixel World, dynamic programming efficiently allows to incorporate real-world constraints and delivers an optimal segmentation with respect to freespace and obstacle information.

Given the left camera image $\mathbb{I}$ of a stereo image pair and the corresponding disparity image $\mathbb{D}$ (all of size $w \times h \in \mathbb{N}^2$), a multi-layered *Stixel World* corresponds to a column-wise segmentation $L \in \mathbb{L}$ of $\mathbb{I}$ into the classes $\mathbb{C} = \{o, g\}$ (*object* and *ground*/*road*) of the following form

$$\begin{aligned} L &= \{L_u\}, \text{ with } 0 \le u < w \\ L_u &= \{s_n\}, \text{ with } 1 \le n \le N_u \le h \\ s_n &= \left\{v_n^b, v_n^t, c_n, f_n(v)\right\}, \text{ with } 0 \le v_n^b \le v_n^t < h\,,\ c_n \in \mathbb{C} \end{aligned} \tag{1}$$

The total number of segments for each column u is given by N_u. The image row coordinates v_n^b (base point) and v_n^t (top point) mark the beginning and end of each segment s_n. Further, $f_n(v)$ is an arbitrary function that computes the disparity (or depth) of that segment at row v (with $v_n^b \le v \le v_n^t$). All segments s_{n-1} and s_n are adjacent such that for each segmentation $L_u \in L \in \mathbb{L}$ of column u the following ordering applies

$$\begin{aligned} 0 &= v_1^b \le v_1^t < \ldots < v_{N_u}^b \le v_{N_u}^t = h-1, \\ &\text{with } v_{n-1}^t + 1 = v_n^b,\ 1 < n \le N_u \end{aligned} \tag{2}$$

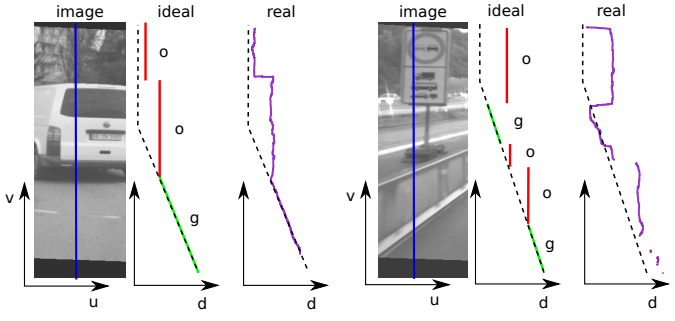

Figure 2: Data model visualization. The blue line across the image marks an exemplary column. Red and green denote the ideal data and segmentation into *object* and *ground*. The dashed line is the expected ground profile. The real disparity measurement vector for the particular scenario is marked purple.

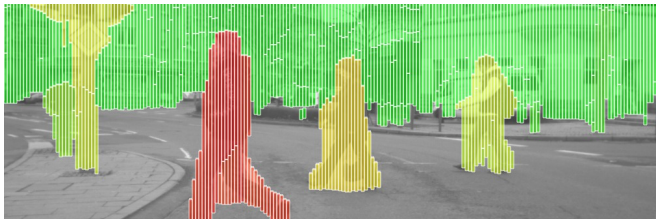

Figure 3: Several pedestrians are crossing our path. Note how accurate their outlines are segmented by the presented approach. Overhanging parts from signs, traffic lights and pedestrians partially violate the ordering and gravity assumption.

Since every segmentation $L \in \mathbb{L}$ conforms to (2) it is implicitly guaranteed that every image point is assigned to exactly one label.

For the labeling step each segment is assigned to either *object* or *ground*. All segments are modeled as piecewise planar surfaces. Thus, the choice for the function f_n is reduced to the set of linear functions. The idea of using linear functions is illustrated in Figure 2.

However, the Stixel result does not solely depend on the measured input data but is regularized by a certain set of physically motivated world assumptions. This includes the following:

- Bayesian information criterion: The number of objects captured along every column is small. Dispensable cuts should be avoided.
- Gravity constraint: Flying objects are unlikely. The ground adjacent object segment should stand on the ground surface.
- Ordering constraint: The upper of two staggered *object* segments usually has a greater depth than the lower one.

Searching for the *Stixel* representation that matches best with the above criteria emerges as a typical MAP estimation problem. Therefore, we search the most probable labeling L^*

$$L^* = \arg\max_{L \in \mathbb{L}} P(L \mid \mathbb{D}) \tag{3}$$

In order to achieve real-time capability, neighboring columns are considered as independent. Hence, L is reduced to the column labeling L_u. As a result we obtain

$$P(L \mid \mathbb{D}) \sim \prod_{u=0}^{w-1} P(D_u \mid L_u) \cdot P(L_u)\,. \tag{4}$$

For our experiments we focus on a stereo vision based evaluation of traffic scenarios. However, in our contribution we present results for both stereo vision data and laser data.

The stereo camera system has a resolution of $1024\,\text{px} \times 440\,\text{px}$, a focal length of $1250\,\text{px}$ and a base length of $22\,\text{cm}$. The used implementation for SGM stereo runs on FPGA hardware at a rate of $25\,\text{Hz}$ with a valid disparity range of $d_{\min} = 0$ to $d_{\max} = 127$.

All further processing is done on the CPU. Thereby, all precomputation for a stereo image pair and a *Stixel* width of $5\,\text{px}$ is done in $5\,\text{ms}$. The solving step via dynamic programming runs within $60\,\text{ms}$. Examples of our approach are illustrated in Figure 3 and Figure 4. This real-time capable approach can also be used to fuse the information of multiple data sources.

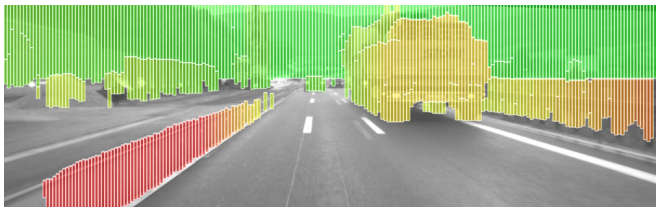

Figure 4: Scenario within a highway construction site containing multiple staggered ground and object segments at different depths. In addition, this scene features a quite far view. The leading car is observed at a distance of $75\,\text{m}$.