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

Abstract

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. Real-time high-level vision systems require a compact and explicit representation of that data which allows for efficient attention control, object detection, and reasoning.

Because man-made environments are dominated by planar horizontal and vertical surfaces we approximate the three dimensional scenery by using sets of thin planar rectangles called Stixels. This medium level representation serves as input for further processing steps and applications. Using this novel representation those are not required to process the large amounts of raw 3D data individually.

This reconstruction is addressed by means of a unified probabilistic approach. Dynamic programming allows to incorporate real-world constraints such as perspective ordering and delivers an optimal segmentation with respect to freespace and obstacle information. We present results for both stereo vision data and laser data. The real-time capable approach can also be used to fuse the information of multiple data sources.

1 Introduction

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, since applications originating from GIS, robotics, or driver assistance often can not afford to evaluate every single depth measurement individually. They ask for a medium level representation that allows structured access to the scene data independent of the particular application without neither being too specific nor too generalizing.

The geometry in man-made environments is dominated by two basic types: Horizontal and vertical planar surfaces, a characteristic increasingly exploited for 3D reconstruction and object modeling. While horizontal surfaces generally correspond to the ground, *i.e.* roads, sidewalks, or soil, the vertical ones relate to objects, such as solid infrastructure, pedestrians, or cars. Due to their inherent orthogonality, these two models render as very distinctive.

Recently, we have presented a medium level representation called the *"Stixel World"* with the objective to efficiently model urban environments with respect to freespace and obstacle information [2]. The proposed *Stixel* computation scheme is a bottom-up approach that cascades multiple independent steps until the final *Stixel* representation is extracted.

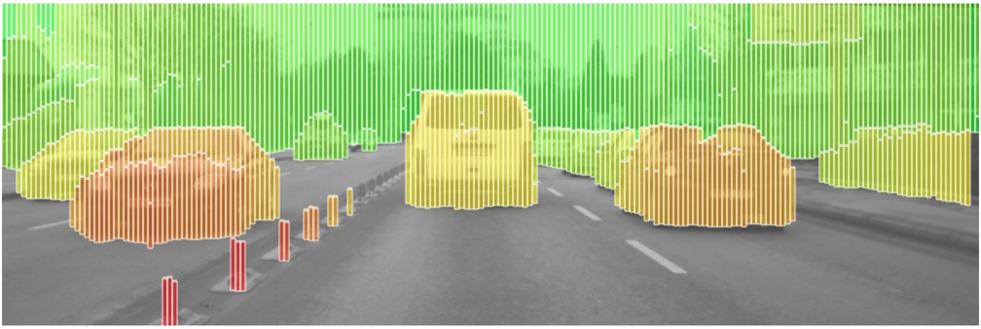

Figure 1: The multi-layer *Stixel World* result as output of the optimization. The captured scene is segmented into planar *Stixel* segments that correspond to either ground or object. The color represents the distance to the obstacle with red being close and green far away. Grey pixels belong to the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

Even though these individual steps use global optimization, the assembled result is not optimal. Thus, we extend that idea to a probabilistic approach that permits the extraction of the Stixel representation by means of a unified global optimal scheme.

In contrast our previous work, objects are allowed to be located at multiple depths in a column. Additionally, distinctive prior knowledge is incorporated into the reconstruction process in order to be able to extract even more detailed information, such as exemplified in Figure 1.

The remainder of this paper is structured as follows: Section 2 briefly points out related work. The *Stixel World* model is discussed and extended accordingly in Section 3 for which Section 4 offers a Bayesian formulation of the optimal *Stixel* generation task. Experimental results are presented in Section 5, while Section 6 concludes this contribution.

2 Related Work

Mapping depth information to local 2D or 3D occupancy grids [5, 18, 24] or digital elevation maps [14] is common practice in order to model the likelihood of the environment to be occupied. Such information is utilized further, *i.e.* for extracting scene attributes such as freespace [1, 12] and obstacle information [20], the location of curbs and sidewalks [19, 22] and various other scene and application relevant features.

A recurring and central key aspect is to explicitly use a priori scene knowledge for 3D reconstruction. The majority of structures in man-made environments consist of piecewise smooth and planar surfaces that have either horizontal or vertical orientation. Such prior knowledge should be incorporated early within the reconstruction process, *e.g.* as done with semi-global matching (SGM) stereo in [10]. SGM works in a dynamic programming [3] (DP) fashion, where slight disparity changes are penalized with rather small and constant costs to prefer the reconstruction of slanted surfaces. A stereo scheme that massively leverages three-dimensional surface planarity is plane sweeping stereo [4]. An extension for that approach has been presented by Gallup *et al.* [7]. The authors incorporate prior knowledge about the location and orientation of planes into the reconstruction process.

Yet exploiting environmental regularities does not end with the plain depth map result. Micusik *et al.* [17] presented a bottom-up approach for extracting 3D structures from panoramic images taken in urban environments. They make extensive use of piecewise pla-

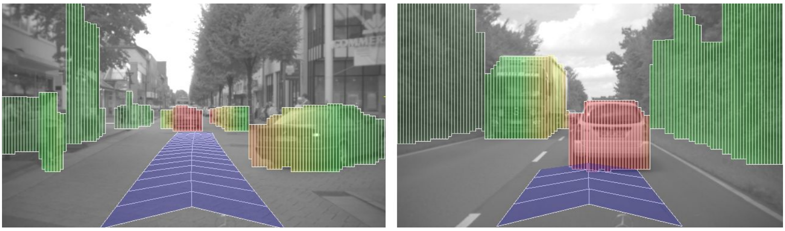

Figure 2: The *Stixel World* as published by Badino *et al.* An urban scenario is shown with *Stixel* approximating cars and infrastructure on the left. The right example shows a rural road scenario.

narity and “L-shape” priors at multiple points in their reconstruction process.

With the objective to efficiently model urban traffic scenarios, we presented a medium level representation called the *“Stixel World”* that unifies freespace and obstacle information in a single and compact scene representation [2]. That representation provides the basis for our work and is discussed separately in Section 3 of this paper.

A similar approach has been published by Gallup *et al.* [8] with the objective to create 3D volumetric object models. Multiple depth maps from different views are accumulated in a single Cartesian histogram-based elevation map. Thereafter, each cell is split into alternating *empty* and *occupied* box volumes. By relying on DP, the authors achieve an optimal segmentation for every cell of the grid.

In [6], Felzenszwalb *et al.* present a probabilistic image segmentation approach that uses appearance to assign semantic information to certain regions of an image. The image is segmented using a continuous upper and lower bound. The resulting upper part is called *background*, the middle region is assigned to *object* and the bottom region to *floor*. The authors rely on DP. However, their approach is limited to one object per column only.

Further, Liu *et al.* [16] use appearance cues to assign semantics to an image using a five parts model (*top*, *left*, *right*, *bottom*, and *center*). Their used model constraints are quite strict and thus inflexible while their graph-cut based approach only approximates a 2D optimum.

In [11], Hoiem *et al.* present a scheme to assign the labels of type *sky*, *vertical* (object) and *planar* (ground) to super-pixels. Therefore, the authors rely on a greedy algorithm while exploiting pairwise patch affinities. They use an appearance-based boosted decision-tree classifier on a trained data set to infer the probabilities for the class affiliation.

3 The Stixel Model

With the goal to efficiently model the content of 3D urban traffic environments [2], Badino *et al.* defined a medium level representation called the *“Stixel World”*. The *Stixel* representation is characterized to be compact, robust to outliers and easy to access.

The space in front of the car is split into two adjacent regions: Horizontal freespace up to the base point of the first obstacles and a set of *Stixels* approximating the obstacle. A single *Stixel* is defined as a thin earthbound rectangle with a fixed pixel width and vertical pose. It is described by just two parameters: A distance and a height value. This representation achieves an enormous reduction of the input data volume of half a million disparity measurements to a few hundred *Stixels* only, while encoding freespace and obstacle information for the whole scenario. An exemplary result for an urban scenario is depicted in Figure 2.

According to that approach, the *Stixel World* is constructed in a cascade of multiple steps: Mapping disparities to occupancy grids, freespace computation, height segmentation, and a final *Stixel* extraction. Such a cascade is prone to errors, *e.g.* missed objects in the freespace computation can not be corrected in subsequent steps. Further, the proposed scheme contains multiple thresholds and nonlinearities (*e.g.* a height constraint when creating the depth map). Only taking into account the first obstacle along every viewing angle can cause to miss relevant objects (*e.g.* a pedestrian standing behind an engine hood).

3.1 Extension of the Stixel World

Hence, our contribution is a probabilistic approach to compute the *Stixel World* for a stereo image pair in a single global optimization step. In addition, we lift the constraint of the *Stixel* to touch the ground surface and allow for multiple *Stixels* along every column of the image, altering the problem of *Stixel* generation into a segmentation problem related to the work of Felzenszwalb *et al.* [6] or Gallup *et al.* [8]. An example for our method is depicted in Figure 1.

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 $\mathbb{C} = \{\mathrm{o}, \mathrm{g}\}$ (*object* and *ground/road*) of the following form

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

The total number of segments for each column u is given by N_u. In this notation, 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 $v_n^b \leq v \leq v_n^t$). 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

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

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

3.2 The Data Model

In order to solve the segmentation task efficiently, we decide to work in image coordinates by using the v-disparity space [15, 21]. This is advantageous for a couple of reasons. Firstly, no extra computation time is required for triangulation or projection. Secondly that coordinate space has inherently finite boundaries, which is beneficial when working with probability densities. Besides that, we do not have to deal with additional quantization artifacts, a common problem of mapping measurements to grids or voxel spaces. Also, the noise characteristic of the depth measuring sensor is preserved and can be considered directly.

For the labeling we decide for each possible segment to be either *object* or *ground*. All segments are modeled as piecewise planar surfaces. Consequently, the choice for the function f_n is reduced to the set of linear functions. Given that the world geometry in man-made environments mainly consists of either vertical or horizontal surfaces, this function set is reduced even further. Therefore, object segments are assumed to have a constant disparity, such that the corresponding function is given as $f_n^o(v) = \mu_n$, where μ_n is the average disparity

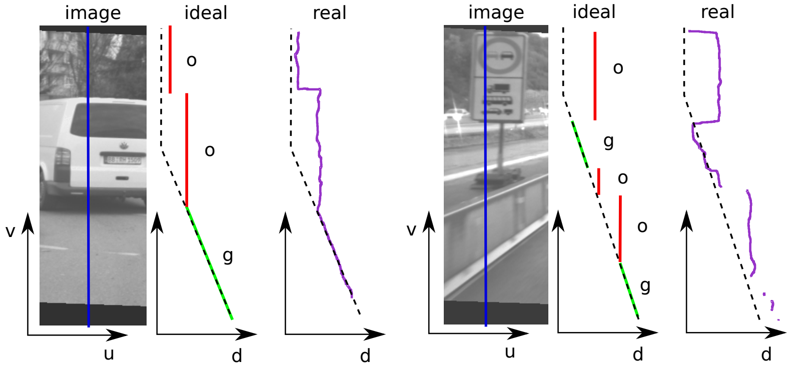

Figure 3: Data model visualization. The blue line across the image marks an exemplary column. Red and green denote the ideal data and segmentation into *object* and *ground*. The dashed line is the expected ground profile. The real disparity measurement vector for the particular scenario is marked purple.

within s_n. The expected ground surface for every *ground* labeled segment is modeled as $f_n^g(v) = \alpha \cdot (v_{\text{hor}} - v)$, where α is the expected ground disparity gradient and v_{hor} is the row coordinate of the horizon. Both parameters are extracted from the known camera geometry. The idea of using these linear functions is illustrated in Figure 3.

4 Stixel Computation as a Probabilistic Approach

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 [8]: 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 is expected to have a greater depth than the lower one. Reconstructing otherwise (*e.g.* for traffic lights, signs or trees) is still possible if sufficiently supported by the input data.

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^*:

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

Applying the Bayes' theorem allows to write the posterior probability $P(L \mid \mathbb{D})$ as

$$P(L \mid \mathbb{D}) \sim P(\mathbb{D} \mid L) \cdot P(L), \tag{4}$$

the product of the conditional probability of $\mathbb{D}$ given L and the prior probability $P(L)$ of L. The neglected normalization factor $P(\mathbb{D})$ is irrelevant when seeking the maximum of the posterior probability. $P(\mathbb{D} \mid L)$ rates the possibility of the input $\mathbb{D}$ given a certain labeling L and thus represents the data term for the optimization. The second term $P(L)$ embodies the overall probability for each individual labeling L and is the lever to model such regularization as listed above.

In order to achieve real-time capability, neighboring columns are considered as independent. Hence, L is reduced to the column labeling L_u. Further, we consider the individual measurements $d_{u,v} \in \mathbb{D}$ as mutually independent and thus also generalize the disparity input to the vertical disparity vector $D_u \in \mathbb{D}$. Additionally, the data within D_u is assumed as independent from all labels $L_{\hat{u}}$ with $u \neq \hat{u}$. As a result we obtain

$$P(L \mid \mathbb{D}) \sim \prod_{u=0}^{w-1} P(D_u \mid L_u) \cdot P(L_u) . \tag{5}$$

4.1 The Conditional Probability Density

The conditional probability density $P(D_u \mid L_u)$ has the objective to rate the likelihood of the input data given a labeling L_u. It is constructed as follows:

$$\begin{aligned} P(D_u \mid L_u) &= \prod_{n=1}^{N_u} \prod_{v=v_n^b}^{v_n^t} P_D(d_v \mid s_n, v) \\ P(d_v \mid s_n, v) &= \begin{cases} P_\mathrm{D}(d_v \mid s_n, v) \cdot (1 - p_{\nexists}^{c_n}) & \text{, if } \exists(v) = 1 \\ p_{\nexists}^{c_n} & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

The data term $P_\mathrm{D}(d_v \mid s_n, v)$ denotes the probability for a disparity measurement $d_v \in D_u$ at column v given the segment s_n. For that purpose, one has to be aware of two facts: Not every pixel has a valid disparity measurement and if it does, it might be an outlier. In order to consider such sensor characteristics, an outlier rate p_out is considered that models the probability to encounter an outlier. Also, a mapping $\exists(v)$ and a probability $p_{\nexists}^{o}$ is defined. $\exists(v)$ equals '1' if the disparity d_v at v is valid (*i.e.* $0 \leq d_v < 128$) and '0' otherwise. $p_{\nexists}^{o}$ is the probability to observe an object given that d_v is invalid. With that in mind, the data term $P_\mathrm{D}(d_v \mid s_n, v)$ is defined as a mixture model that consists of a uniform distribution to model the chance to encounter outliers and a Gaussian distribution to rate the affinity of d_v to s_n by

$$P_\mathrm{D}(d_v \mid s_n, v) = \frac{p_\mathrm{out}}{d_\mathrm{max} - d_\mathrm{min}} + \frac{\neg p_\mathrm{out}}{A_\mathrm{norm}} \cdot \frac{1}{\sigma^{c_n}(d_v, v) \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{d_v - f_n(v)}{\sigma^{c_n}(f_n, v)}\right)^2} \tag{7}$$

The Gaussian parameter $\sigma^{c_n}(d_v, v)$ incorporates the noise model for the disparity measurement. It depends on the segment type c_n, the expected disparity noise, the camera geometry and tilt angle accuracy. It also considers by what degree objects are allowed to violate the constant disparity assumption. A_norm is a normalization term computed with the integral

$$A_\mathrm{norm} = 0.5 \cdot \left(\mathrm{erf}\left(\frac{d_\mathrm{max} - f_n(v)}{\sqrt{2} \cdot \sigma^{c_n}(f_n, v)} \right) - \mathrm{erf}\left(\frac{d_\mathrm{min} - f_n(v)}{\sqrt{2} \cdot \sigma^{c_n}(f_n, v)} \right) \right) . \tag{8}$$

4.2 Modeling A Priori Knowledge

The second term $P(L_u)$ of equation (5) models our world model expectation and a-priori knowledge. In contrast to the conditional probability density $P(D_u|L_u)$ it does not contain any dependencies to the input data. Instead that term considers semantic aspects as listed above. For this purpose, pairwise mutual dependencies are modeled between all adjacent segments s_{n-1} and s_n. $P(L_u)$ is derived by

$$P(L_u) = P(s_1, \ldots, s_{N_u}) = P(s_1) \cdot \prod_{n=2}^{N_u} P(s_n \mid s_{n-1}), \tag{9}$$

where the former term $P(s_1)$ states the probability of a particular occurrence of the first segment $s_1 = \{v_1^b, v_1^t, c_1, f_1(v)\}$. The second part incorporates semantic aspects between adjacent segments, such as ordering or gravity regularization. $P(s_1)$ separates as follows:

$$\begin{aligned} P(s_1) &= P\left(v_1^b, v_1^t, c_1, f_1(v)\right) \\ &= P\left(v_1^b\right) \cdot P\left(v_1^t\right) \cdot P\left(c_1 \mid v_1^t\right) \cdot P\left(f_1(v) \mid c_1\right) \end{aligned} \tag{10}$$

Hereby $P\left(v_1^b\right)$ is straightforward to compute: According to equation (2), v_1^b must equal to 0. Further, we obtain:

$$P\left(v_1^b\right) = \begin{cases} 1 & ,v_1^b = 0 \\ 0 & \text{otherwise} \end{cases} \qquad P\left(c_1 \mid v_1^t\right) = \begin{cases} 1 & ,v_1^t > v_{\text{hor}},\, c_1 = \text{o} \\ 0 & ,v_1^t > v_{\text{hor}},\, c_1 = \text{g} \\ 0.5 & \text{otherwise} \end{cases}$$

$$P\left(v_1^t\right) = {}^1\!/_{h-1} \qquad P\left(f_1(v) \mid c_1\right) = \begin{cases} 1 & ,c_1 = \text{o}, f_1(v) = \mu_1 \\ 1 & ,c_1 = \text{g}, f_1(v) = \alpha \cdot (v_{\text{hor}} - v) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The conditional density term $P(s_n \mid s_{n-1})$ is determined by separating $P(s_n \mid s_{n-1}) = P\left(v_n^b, v_n^t, c_n, f_n(v) \mid v_{n-1}^b, v_{n-1}^t, c_{n-1}, f_{n-1}(v)\right)$. This appears sophisticated, but is simplified to a few factors:

$$\begin{aligned} P(s_n \mid s_{n-1}) &= P\left(v_n^b \mid v_{n-1}^t\right) \cdot P\left(v_n^t \mid v_{n-1}^t\right) \cdot P\left(c_n \mid v_{n-1}^t, c_{n-1}\right) \\ &\cdot\ P\left(f_n(v) \mid c_n, v_{n-1}^t, c_{n-1}, f_{n-1}(v)\right) \end{aligned} \tag{12}$$

In contrast to the last two terms, the conditional probabilities $P\left(v_n^b \mid v_{n-1}^t\right)$ and $P\left(v_n^t \mid v_{n-1}^t\right)$ lack a deeper meaning and plainly express that the value range for v_n^b and v_n^t is limited by v_{n-1}^t. They are determined by

$$P\left(v_n^b \mid v_{n-1}^t\right) = \begin{cases} 1 & ,\, v_n^b = v_{n-1}^t + 1 \\ 0 & \text{otherwise} \end{cases}, \text{and } P\left(v_n^t \mid v_{n-1}^t\right) = \begin{cases} {}^1\!/_{h - v_{n-1}^t - 2} & ,\, v_n^t > v_{n-1}^t \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Just like $P(c_1 \mid v_1^t)$, the term $P\left(c_n \mid v_{n-1}^t, c_{n-1}\right)$ expresses not to expect street occurrence above the horizon. It also models the expectation of adjacent segments to occur in a certain order. For instance, it is rather unlikely to observe ground occurrence behind an object. $P\left(c_n \mid v_{n-1}^t, c_{n-1}\right)$ is modeled using a look-up-table that is omitted due to lack of space.

The last term $P\left(f_n(v) \mid c_n, v_{n-1}^t, c_{n-1}, f_{n-1}(v)\right)$ models two more aspects: The probability for floating objects (ordering constraint and gravity constraint) and the probability to have objects below the ground surface. It is defined by Table 1. p_{ord} corresponds to the ordering regularization and models the probability of two staggered objects to violate the ordering assumption, such that s_n has a larger disparity and thus is closer than s_{n-1}. The second variable p_{grav} models the probability of ground adjacent objects to hover and hence not to touch

c_n	c_{n-1}	condition	P
o	o	$\mu_n > \mu_{n-1} + \Delta_d(\mu_{n-1}, \Delta_Z)$	$p_{\text{ord}}/\mu_{n-1}-\Delta_d$
o	o	$\mu_n \leq \mu_{n-1} - \Delta_d(\mu_{n-1}, \Delta_Z)$	$1-p_{\text{ord}}/d_{\max}-\mu_{n-1}-\Delta_d$
o	o	$\lvert\mu_n - \mu_{n-1}\rvert < 2 \cdot \Delta_d(\mu_{n-1}, \Delta_Z)$	0
o	g	$\mu_n > f_{n-1}(v^t_{n-1}) + \varepsilon$	$p_{\text{grav}}/d_{\max}-f_{n-1}(v^t_{n-1})-\varepsilon$
o	g	$\mu_n < f_{n-1}(v^t_{n-1}) - \varepsilon$	$p_{\text{blg}}/f_{n-1}(v^t_{n-1})-\varepsilon$
o	g	$\lvert\mu_n - f_{n-1}(v^t_{n-1})\rvert < 2 \cdot \varepsilon$	$1-p_{\text{grav}}-p_{\text{blg}}/2\cdot\varepsilon$
g	o$\vee$g	$f_n(v) = \alpha \cdot (v - v_{\text{hor}})$	1, otherwise 0

Table 1: Look-up table for $P(f_n(v)|c_n, v^t_{n-1}, c_{n-1}, f_{n-1}(v))$ that models the probability for function $f_n(v)$ given a configuration c_n, c_{n-1}, d_{n-1} and v^t_{n-1}. μ_n is the mean disparity of segment s_n with $c_n =$ object.

the ground surface. Object segments s_n and s_m are not allowed to coexist within a certain distance $\pm\Delta_Z$. That range is mapped to disparities by $\Delta_d(\mu, \Delta_Z)$.

The third and last variable p_{blg} denotes the probability for objects to have a base point below the ground. In this notation, $f_{n-1}\left(v^t_{n-1}\right)$ with $c_{n-1} = \text{g}$ is the end disparity of the ground segment. For our results we choose $p_{\text{ord}} = p_{\text{grav}} = 0.1$ and $p_{\text{blg}} = 0.001$, which renders especially the last configuration as very unlikely. The parameter ε denotes the range in which violations of the gravity and ground assumption are tolerated. These parameters are chosen as $\Delta_Z = 1.5\,\text{m}$ and $\varepsilon = 1.5\,\text{px}$ disparities.

4.3 Solving for L^* with Dynamic Programming

Dynamic Programming (DP) [3] has been successfully applied as a solving scheme for a vast number of optimization problems. It has the major benefit of yielding the global optimum of a discrete problem that exhibits optimal substructure non-iteratively in polynomial time.

Relying on DP is what makes solving for L^* computable in real-time. We seek a labeling for a stereo image pair (each of size $w \times h$) and allow for a label to be set individually at every pixel of every column. Modeling mutual dependencies for adjacent labels results in a computational complexity of $\mathcal{O}\left(n^3\right) = w \times h^2/2 \times |\mathbb{C}|$ using the Landau notation [23].

All data terms (see section 4.1) required within the solving step can be precomputed using the log-likelihood [13] of their probabilities and thus do not impact the run-time performance of the optimizer. The a-priori terms (see section 4.2) are also either precomputed or are evaluated in place during the optimization.

5 Experimental Results

For our experiments we focus on a stereo vision based evaluation of traffic scenarios. 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 $22\,\text{cm}$. It is mounted behind the windshield at a height of $h_{\text{cam}} = 1.17\,\text{m}$ and with a downwards tilted angle $\alpha_{\text{cam}} = 0.063\,\text{rad}$.

The used implementation for SGM stereo runs on FPGA hardware at a rate of 25 Hz with a valid disparity range of $d_{\min} = 0$ to $d_{\max} = 127$ and an assumed uncertainty of $\sigma_d = 0.4\,\text{px}$. For the disparity outlier rate we assume $p_{\text{out}} = 0.1$.

All further processing is done on the CPU (Core-i7 980X, $6 \times 3.4\,\text{Ghz}$, 6 GB of RAM). Thereby, all precomputation for a stereo image pair and a *Stixel* width of 5 px is done in

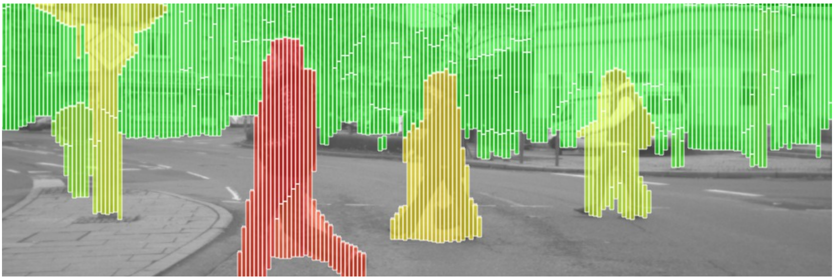

Figure 4: Several pedestrians are crossing our path. Note how accurate their outlines are segmented by the presented approach. Overhanging parts from signs, traffic lights and pedestrians partially violate the ordering and gravity assumption.

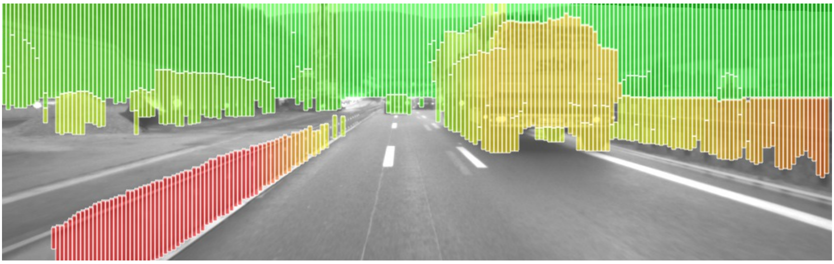

Figure 5: Scenario within a highway construction site containing multiple staggered ground and object segments at different depths. In addition, this scene features a quite far view. The leading car is observed at a distance of 75 m.

5 ms. The solving step via dynamic programming runs within 60 ms. Due to the smoothing characteristic of SGM scaling down the image by a factor of two comes without noteworthy impact to the quality of the scene reconstruction. However, this reduces the computation time significantly, such that solving is done in real-time within 15 ms. Yet the precomputation remains unchanged, because all look-up tables are still computed at full resolution in order to minimize the potential loss of accuracy as a result from scaling.

The depicted scenarios feature various types of objects and different scene constellations. An exemplary urban scenario with pylons, cars and solid infrastructure is given in Figure 1. Figure 4 shows an multiple pedestrians and objects that partially violate the ordering and gravity assumptions. Figure 5 illustrates a highway scenario with guardian rails, an opposing

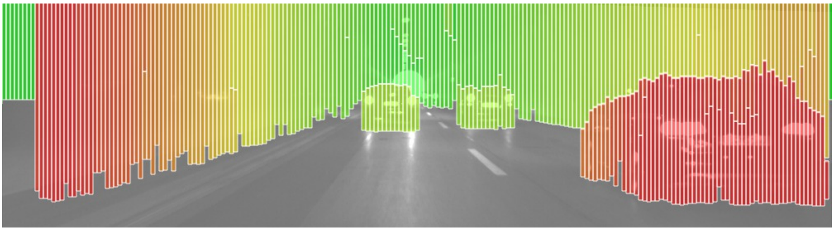

Figure 6: This scene was captured in a highway tunnel and features three leading cars ahead of us. Poor lighting conditions, strong light reflectance on the road surfaces. Altogether, this scenario is considered quite challenging.

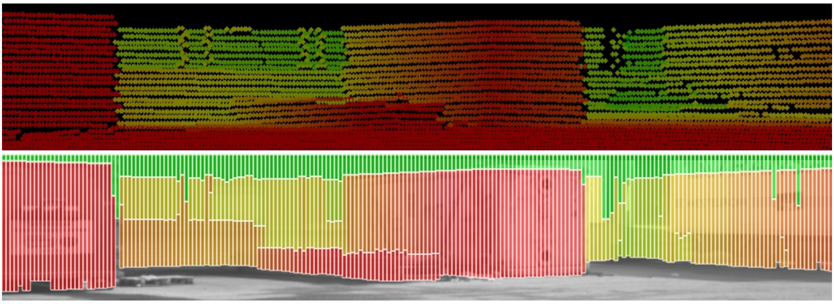

Figure 7: The *Stixel World* created from sparse 3D point cloud measurements obtained with a Velodyne HDL-64E. The upper part shows the LIDAR points, the lower part shows the *Stixel* result.

lane and quite far sight. Figure 6 shows a tunnel scenario that is quite challenging for reasons of poor lighting and a reflecting road surface. All results have been computed with identical parametrization of the algorithm.

The approach is not limited to processing stereo vision data. Therefore, our last example shows *Stixels* created from the 3D data delivered by a Velodyne HDL-64E S2 [9] LIDAR. Figure 7 shows the sparse LIDAR points and the resulting *Stixel* representation that has been projected into the image of a camera that has been calibrated relatively to the LIDAR.

6 Conclusions and Outlook

This paper proposes to model real-world scenes by means of a multi-layer *Stixel World*. We tackle the inherent segmentation task as a MAP-problem. This task has been formulated such that the optimization can be solved efficiently by means of DP.

Seeking the most probable interpretation results in a highly robust approach as shown by our examples. The overall algorithm has proven as quite parameter-insensitive. Thus, scenario dependent fine tuning is only required for extreme weather conditions, when sensor characteristics change drastically.

The presented approach is applicable to process 3D data from other sensors as well. Exemplary results for a modern laser scanner have been presented. Consequently, the proposed approach can also act as a fusion scheme for multiple data sources.

An aspect we did not target was to consider horizontal smoothness, such as Felzenszwalb [6] or Badino [2] did within their DP steps. Yet we argue that this property is partially realized by the smoothing characteristic of SGM. By looking at the results we do not make out a real benefit by enforcing that additionally. Besides that, doing so would turn this segmentation task into a two-dimensional NP-hard labeling problem.

Moreover, appearance cues or temporal coherence such as optical flow or tracking was not used. Without doubt, such information carries the potential to improve the system performance even further if incorporated appropriately. For our purpose, we will focus on an extension of supported classes and will pay dedicated attention to a more diligent treatment of missing or faulty disparities especially in sky and ground regions with low textural information.

References

[1] Hernàn Badino, Uwe Franke, and Rolf Mester. Free space computation using stochastic occupancy grids and dynamic programming. In *Workshop on Dynamical Vision, ICCV*, Rio de Janeiro, Brazil, October 2007.

[2] Hernàn Badino, Uwe Franke, and David Pfeiffer. The stixel world - a compact medium level representation of the 3d-world. In *German Association for Pattern Recognition (DAGM)*, pages 51–60, Jena, Germany, September 2009.

[3] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[4] Robert Collins. A space-sweep approach to true multi-image matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 358–363, June 1996.

[5] Alberto E. Elfes. Sonar-based real-world mapping and navigation. *Journal of Robotics and Automation*, 3(3):249–265, June 1987.

[6] Pedro F. Felzenszwalb and Olga Veksler. Tiered scene labeling with dynamic programming. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3104, San Francisco, CA, USA, June 2010.

[7] David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[8] David Gallup, Marc Pollefeys, and Jan-Michael Frahm. 3d reconstruction using an n-layer heightmap. In *German Association for Pattern Recognition (DAGM)*, pages 1–10, Darmstadt, Germany, September 2010.

[9] Velodyne Headquarters. High definition lidar hdl-64e s2, February 2010.

[10] Heiko Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 807–814, 2005.

[11] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *International Conference on Computer Vision (ICCV)*, pages 654–661, 2005.

[12] Florian Homm, Nico Kaempchen, Jeff Ota, and Darius Burschka. Efficient occupancy grid computation on the gpu with lidar and radar for road boundary detection. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1006–1013, San Diego, CA, USA, June 2010.

[13] Steven M. Kay. *Fundamentals of statistical signal processing: estimation theory*. Number v. 2 in Prentice Hall signal processing series. PTR Prentice-Hall, 1993. ISBN 9780133457117.

[14] In So Kweon and Takeo Kanade. High-resolution terrain map from multiple sensor data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 14: 278–292, 1992.

[15] Raphael Labayrade, Didier Aubert, and Jean-Philippe Tarel. Real time obstacle detection in stereovision on non flat road geometry through v-disparity representation. In *IEEE Intelligent Vehicles Symposium (IV)*, 2002.

[16] Xiaoqing Liu, Olga Veksler, and Jagath Samarabandu. Order-preserving moves for graph-cut-based optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(7):1182–1196, 2010.

[17] Branislav Micusik and Jana Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 0:2906–2912, 2009.

[18] Hans P. Moravec. Robot spatial perception by stereoscopic vision and 3d evidence grids. Technical Report CMU-RI-TR-96-34, Carnegie Mellon University, 1996.

[19] Florin Oniga and Sergiu Nedevschi. Curb detection for driving assistance systems: A cubic spline-based approach. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 945–950, Baden-Baden, Germany, June 2011.

[20] Florin Oniga, Sergiu Nedevschi, Marc-Michael Meinecke, and Thanh Binh To. Road surface and obstacle detection based on elevation maps from dense stereo. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*, Seattle, WA, USA, September 2007.

[21] Mathias Perrollaz, Raphaël Labayrade, Romain Gallen, and Didier Aubert. A three resolution framework for reliable road obstacle detection using stereovision. In *IAPR Conference on Machine Vision Applications (MVA)*, pages 469–472, 2007.

[22] Jan Siegemund, David Pfeiffer, Uwe Franke, and Wolfgang Förstner. Curb reconstruction using conditional random fields. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 203–210, San Diego, CA, USA, June 2010.

[23] Michael Sipser. *Introduction to the Theory of Computation*. International Thomson Publishing, 1st edition, 1996. ISBN 053494728X.

[24] K. M. Wurm, A. Hornung, M. Bennewitz, C. Stachniss, and W. Burgard. OctoMap: A probabilistic, flexible, and compact 3D map representation for robotic systems. In *Proceedings of the ICRA 2010 Workshop on Best Practice in 3D Perception and Modeling for Mobile Manipulation*, Anchorage, AK, USA, May 2010. Software available at http://octomap.sf.net/.