Exploiting the Power of Stereo Confidences

David Pfeiffer
Daimler AG
Sindelfingen, Germany
david.pfeiffer@daimler.com

Stefan Gehrig
Daimler AG
Sindelfingen, Germany
stefan.gehrig@daimler.com

Nicolai Schneider
IT-Designers GmbH
Esslingen, Germany
stz.schneider@daimler.com

Abstract

Applications based on stereo vision are becoming increasingly common, ranging from gaming over robotics to driver assistance. While stereo algorithms have been investigated heavily both on the pixel and the application level, far less attention has been dedicated to the use of stereo confidence cues. Mostly, a threshold is applied to the confidence values for further processing, which is essentially a sparsified disparity map. This is straightforward but it does not take full advantage of the available information.

In this paper, we make full use of the stereo confidence cues by propagating all confidence values along with the measured disparities in a Bayesian manner. Before using this information, a mapping from confidence values to disparity outlier probability rate is performed based on gathered disparity statistics from labeled video data.

We present an extension of the so called Stixel World, a generic 3D intermediate representation that can serve as input for many of the applications mentioned above. This scheme is modified to directly exploit stereo confidence cues in the underlying sensor model during a maximum a posteriori estimation process.

The effectiveness of this step is verified in an in-depth evaluation on a large real-world traffic data base of which parts are made publicly available. We show that using stereo confidence cues allows both reducing the number of false object detections by a factor of six while keeping the detection rate at a near constant level.

1. Introduction

Stereo vision has been an actively researched area for decades. In recent years, stereo algorithms and applications have matured significantly spawning products in fields ranging from industrial automation over gaming up to driver assistance systems. The underlying stereo algorithms and their properties are well understood, at least for the current real-time algorithms, typically approaches based on correlation [20] or semi-global matching (SGM) [10]. Benchmarks that compare stereo algorithms on a 100 % density level are available [19], also for the automotive domain [8].

Figure 1: The Stixel World computed from stereo data. The scene is segmented into free space and vertical obstacles. The color (from red to green) represents the object distance.

The computation of stereo confidences has only recently been researched in more detail. Hu and Mordohai [12] performed an excellent review of known stereo confidence metrics comparing them to ground truth scenes on a pixel level. In related work on confidence estimation for stereo or optical flow computation, the so called sparsification plots are established as the main method to show the effectiveness of the considered confidence metric. This procedure gives a good impression with respect to how well the confidence helps reducing the average error of the disparity map when the least confident values are removed. However, no explicit use of both the disparity map and the confidence map in further processing has been reported so far.

Our work is centered around the driver assistance scenario. The main objective is to robustly extract free space and obstacle information from dense disparity maps and to represent the results in a compact and simple fashion.

The Stixel World, firstly introduced by Badino *et al.* [2], is a very suitable representation for this task. Based on an occupancy map [1, 5], this scheme allows to extract the closest row of objects for each image column. In a generalization of this work, we introduced the multi-layered Stixel World [17] that allows to detect all objects in a scene. A result of this scheme is shown in Figure 1.

This paper extends our Bayesian approach [17] to use stereo confidence cues. The idea is that each disparity mea-

surement is given an individual probability to be an outlier. This probability is inferred by using a data base with annotated video data of different weather and lighting scenarios. Then, the resulting information is directly taken into account in the underlying sensor model during the maximum a posteriori (MAP) estimation. The effectiveness of this procedure is evaluated on a large sequence data base containing different adverse scenarios for the stereo sensor setup. To round things off, the performance is also compared against the straightforward way of using sparsification on the disparity map.

The main contribution of this paper is the first-time fully probabilistic usage of stereo confidences along with the disparity map. Moreover, we introduce modified stereo confidence metrics suited for global stereo algorithms, and link confidence values to disparity outlier probabilities.

The paper is organized as follows: Section 2 describes related work to the field of stereo confidence estimation. We limit ourselves to references that inspired our confidence metrics. In addition, work that makes use of stereo confidences in subsequent processing is analyzed. Besides stereo confidence we also review work on 3D intermediate representations. Section 3 encourages our selection of stereo confidence metrics and their modifications for our application. The resulting confidence values are mapped to outlier probabilities which is described in Section 4. In Section 5, the Stixel World is introduced, followed by the extension to use stereo confidences, also for further applications, in the subsequent Sections 6 and 7. Evaluation on large data sets were conducted for which the results are shown in Section 8. We close this paper with conclusions and an outlook.

2. Related Work

Stereo confidence computation has recently attracted rising attention [4, 9, 12]. So far, most work on stereo confidences focused on local stereo approaches. Haeusler *et al*. [9] applied some of these confidence metrics to SGM. In [12], Hu and Mordohai provide an extensive review of existing stereo confidence metrics, again using local correlation as the underlying stereo method. Their results are obtained by analyzing sparsification measures. To this end, the disparity values are sorted according to their confidence values. Subsequently, those depth measurements with the lowest confidence are dropped and a new error metric is calculated for the remaining pixels.

All applications where a left-right consistency check is applied, *e.g*. [22], consider stereo confidences implicitly. Inconsistent matches are ruled out and thus are not considered any further. Milella and Siegwart [15] explicitly compute stereo confidence and eliminate less confident matches for the use in an iterated closest point (ICP) algorithm for ego-motion estimation. Zhang *et al*. [26] also compute stereo confidence and eliminate less reliable matches by thresholding on the confidence. In addition, the confidence value is used as a weight in plane fitting for 3D reconstruction. The stereo uncertainty (*i.e*. the precision of a stereo measurement) has been incorporated several times into occupancy grid approaches where obstacles are mapped onto a grid structure (*e.g*. [1, 5, 23]).

To describe the relevant information of the scene (free space and obstacles) in a compact fashion, we rely on the Stixel World. The first work on this medium-level representation has been conducted by Badino *et al*. [2]. This computation scheme consists of different, independent processing steps including mapping disparities to an occupancy grid, a free space computation, a height segmentation and the final Stixel extraction. A related, yet more run-time optimized approach has been presented by Benenson *et al*. [3].

With the goal to minimize the total number of individual processing steps, we have presented a method to compute the Stixel World in a global optimization using a probabilistic framework [17]. Also, this work extends the capability of the Stixel World in order to describe arbitrarily staggered scenes with more than one object per image column.

Gallup *et al*. [6] published a probabilistic method for segmenting an n-layer height-map (which is basically a three-dimensional occupancy grid) into box volumes with an alternating state of either *"empty space"* or *"occupied"*.

While particularly the work of Gallup *et al*. and our approach make explicit use of a detailed sensor model by precisely taking the measurement noise and outlier characteristic of the particular sensor setup into account, till now, no approach has taken advantage of stereo confidence cues.

3. Stereo Confidence Metrics

The stereo confidence metrics introduced in [12] have been investigated in conjunction with a local stereo method. We use semi-global matching [10] for our work since it robustly provides real-time and dense disparity estimates while being almost as computationally efficient as local methods [7]. Further, we employ the Census metric as investigated by [11]. For all considered metrics, the cost term from the local method is replaced with the accumlated costs of SGM. Those are obtained by traversing multiple paths (in our case 8). In addition, we perform the well-known left-right consistency (LRC) check that was shown to be very effective for high stereo densities [9]. Yet, we apply it less stringently such that only 5 % of the pixels are removed in typical traffic scenes. Note that an approximate of the LRC is very efficient to compute [16].

In total, we focus on three confidence metrics. First of all, we choose two of the most promising metrics from [12], namely peak-ratio naive (PKRN) and maximum likelihood metric (MLM). Both metrics performed clearly above average both indoors and outdoors.

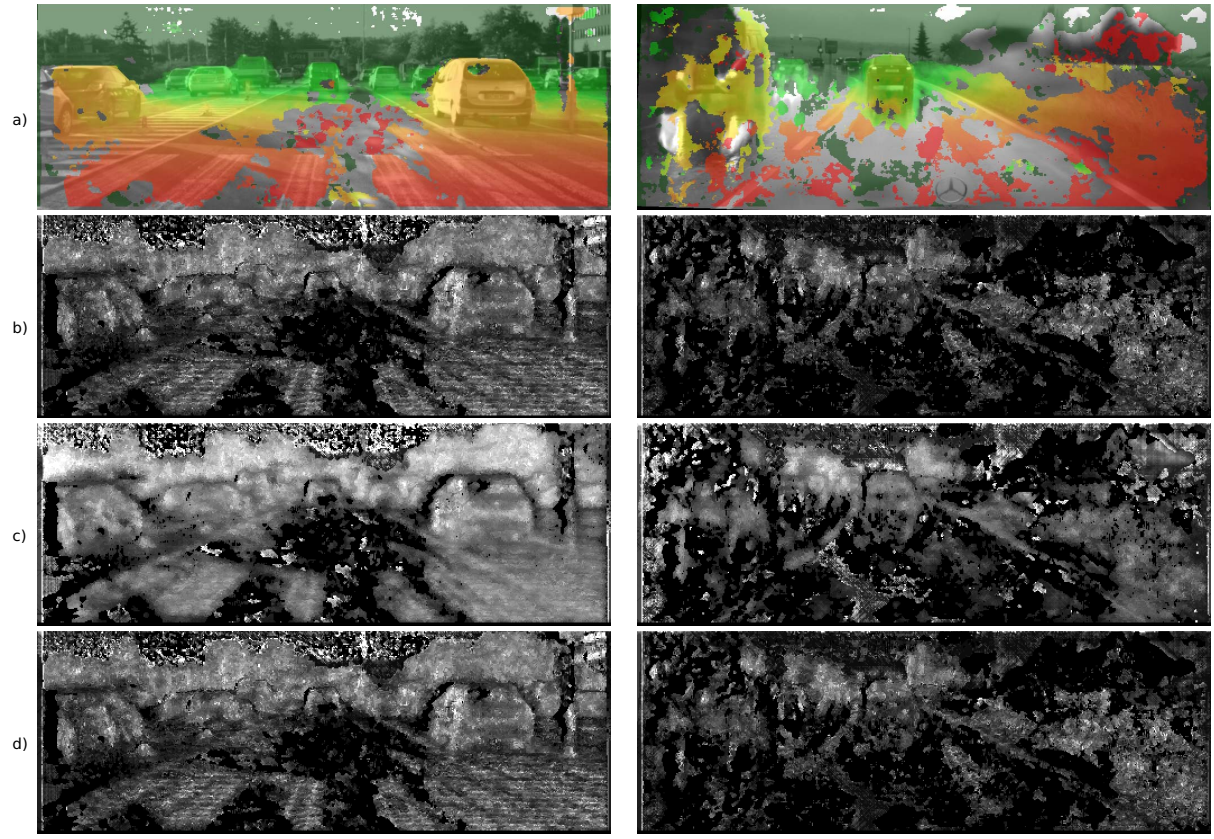

Figure 2: Two challenging scenarios for stereo computation due to disturbances caused by either strong reflections in the windshield (left side) and heavy rain (right side). The upper row a) shows the partially erroneous SGM-based depth map. The color encodes red for close and green for far away. The lower rows visualize the particular stereo confidence cues which is b) LC, c) PKRN, and d) MLM. The brighter a pixel is, the higher is the confidence that the depth measurement is correct.

The PKRN and MLM metric are modified to become:

$$PKRN = \frac{C_2 + \epsilon}{C_1 + \epsilon} - 1 \text{ and} \tag{1}$$

$$MLM = \frac{\mathrm{e}^{-C_1/2\sigma^2}}{\sum \mathrm{e}^{-C_i/2\sigma^2}}, \tag{2}$$

where C_1 is the cost minimum and C_2 is the second smallest cost. We exclude very similar, adjacent costs to not penalize disparity results around half integer values. For MLM, σ represents the disparity uncertainty which is chosen conservatively high (*i.e.* $\sigma = 8$), also for a more uniform distribution. Although the PKRN modifications slightly violate the confidence ordering of the original metric, they have the following advantages over the original counterpart:

- The rare case of a singularity with a denominator of zero is avoided.
- Small changes in costs due to noise at low cost levels do not impact the resulting metric heavily.
- By choosing $\epsilon = 128$, the dynamic range of this metric is limited and the distribution is rather uniform between zero and one for typical scenes.

As the third metric the local curve (LC) information [24] of the equiangular fit is used. It is very similar to the curvature fit of parabola interpolation schemes [20] and it comes with no additional computation as it is a by-product of the subpixel interpolation step. LC is computed as

$$LC = \frac{max(C_+, C_-) - C_{\text{min}}}{\gamma}. \tag{3}$$

The costs C_+ and C_- are adjacent to the optimal disparity C_{min}. To obtain a nicely spread distribution, we choose $\gamma = 480$.

The results of both the stereo matching and the confidence metrics are illustrated in Figure 2. For this purpose, two different situations are shown that pose quite a challenge for the stereo matching. The first one exhibits strong textural patterns caused by scattered sunlight in the windshield which clearly misleads the stereo estimation. The second features heavy rain which results in a blurred vision and strong reflections on the road surface.

For practical reasons, all confidence metrics are scaled and bound to the interval $[0 \ldots 1]$. However, they are not to be mistaken as a probabilistic measure.

4. Using Confidences as Outlier Probabilities

Before turning to the core topic of this section it is important to establish a common conception of confidence metrics as discussed in this work.

Instead of simply computing the disparity measurement d for a pixel, we assume the used stereo scheme to output pairs of values (d, c), with $d \in \mathbb{D}$ and $c \in [0, 1]$. The term $\mathbb{D}$ denotes the set of valid disparity values, *i.e.* $\mathbb{D} = [0, 127]$, and c is the corresponding confidence value of d. According to their construction in the previous section, the interpretation of c is as follows: $c \to 1$ if the measured disparity is rated as *"rather good"* and $c \to 0$ in case it is rated *"rather bad"*, *c.f.* Figure 2.

It is to be expected that this value c strongly depends on various aspects: Foremost the confidence metric itself (*i.e.* PKRN, LC, or MLM), the used stereo scheme (in our case SGM), and the corresponding parameter choice for that particular stereo scheme.

Since we intend to use confidences in a probabilistic framework for Stixel computation, a mapping from the particular confidence metric to an actual outlier probability is required. This mapping $p(c) \to p \in [0, 1]$ is constructed in a way that $p(c) = p(\mathrm{o} \mid c) = p(\text{"}d \text{ is an outlier"} \mid c)$.

Thus, the central question is how this mapping $p(\mathrm{o} \mid c)$ is obtained. A possible method for inference is to use ground truth data from rendered scenes. Such a mapping was implicitly obtained for optical flow estimates in the work of Mac Aodha *et al.* [14]. However, it has been shown that artificially creating realistic ground truth sensor data (as it would be obtained in adverse outdoor environments) is an utterly complex and rather unsolved challenge, *c.f.* [13, 19, 20].

For this reason, a training-based approach is used. It works with the same type of sensor data and stereo algorithm that we later run our vision algorithms on. Similar to using the ground truth data, the underlying idea is straightforward: a human inspector annotates regions in the stereo map using the binary labels *"inlier"* and *"outlier"*. The resulting data set is used to infer the proper mapping function $p(\mathrm{o} \mid c)$. Note that this step has to be done only once per stereo scheme (or stereo parameter choice) but is independent of the used confidence metric. Subsequently, $p(\mathrm{o} \mid c)$ is obtained:

$$p(\mathrm{o} \mid c) = \frac{p(c \mid \mathrm{o}) \cdot p(\mathrm{o})}{p(c)} = \frac{p(c \mid \mathrm{o}) \cdot p(\mathrm{o})}{p(c \mid \mathrm{o}) \cdot p(\mathrm{o}) + p(c \mid \mathrm{i}) \cdot (1 - p(\mathrm{o}))} \tag{4}$$

At first sight this might seem trivial, but it puts us in a quite comfortable position: By using the labeled data it is straightforward to extract $p(c \mid \mathrm{o})$ as well as $p(c \mid \mathrm{i})$. The parameter $p(\mathrm{o})$ can be obtained the same way. Alternatively, this term can also be taken as an additional parameter to gain influence on the performance of the subsequently used algorithms. Thus, since we seek for maximum robustness of our vision system, a comparatively high but safe outlier rate of $p(\mathrm{o}) = 0.4$ is chosen.

When applying this procedure to the labeled training data set that contains about 40 scenarios featuring different weather and lighting conditions, the results shown in Figure 3 are obtained. Hereby, the MLM metric appears to separate best into inliers and outliers. This impression is confirmed when looking at the overlap of the inlier and outlier distribution: LC yields 55.3 %, PKRN has 52.3 %, and MLM achieves the best result with only 40.5 % overlap.

The right side of Figure 3 shows the obtained confidence mapping $p(\mathrm{o} \mid c)$. In accordance with the separation plots, the MLM curve shows the sharpest distinction.

5. The Stixel World

Modern stereo-based vision systems, like those deployed in the field of driver assistance and many robotic applications, feature an increasing complexity. This refers to both the number of executed vision tasks and the large amount of measurement data that has to be processed in real-time. In most cases, those systems have to comply with a shear number of external requirements, *e.g.* limited CPU-power, low memory, a small I/O-bandwidth, or the crucial need for energy efficiency [21, 25].

In order to tackle this challenge, we have proposed the multi-layered Stixel World [17] that provides the most task relevant scene information from large amounts of three-dimensional point-cloud data in an utterly compact, robust, and easy-to-use medium-level representation.

The basic idea for the Stixel World is that man-made environments are dominated by either horizontal or vertical planar surfaces. While horizontal surfaces typically correspond to the ground, the vertical ones relate to objects, for instance solid infrastructure, pedestrians, or cars.

Following this model of perception, each Stixel approx-

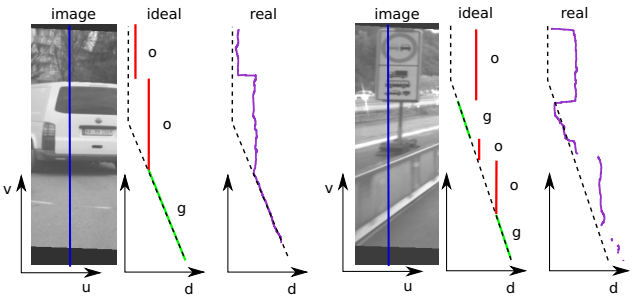

Figure 4: The blue line marks the column for segmentation. Red and green denote the ideal segmentation into object and ground. The dashed line is the expected ground profile and the disparity measurement vector is marked using purple.

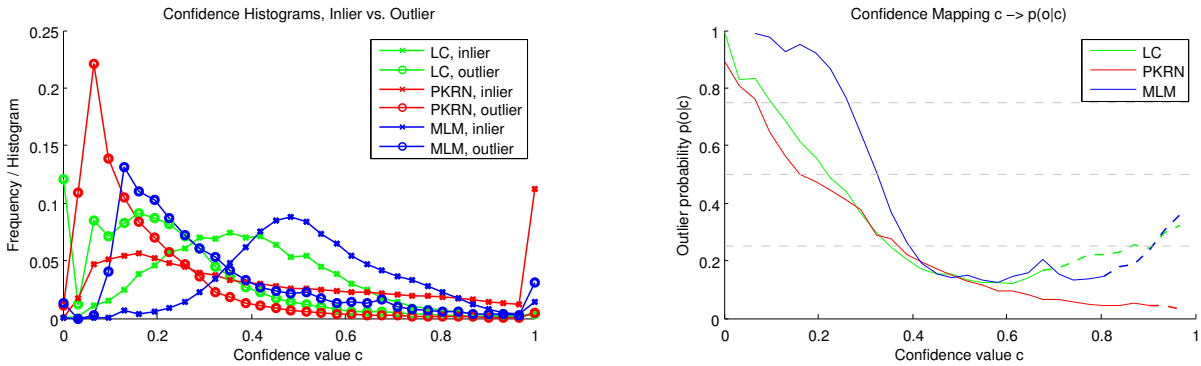

Figure 3: The left figure shows the confidence distribution with respect to the manual labeling of the disparity values (training data) into *"inlier"* and *"outlier"*. All three metrics spread well across the considered interval of $[0 \dots 1]$. According to the visual impression, the separation of inliers and outliers seems to be achieved best by the MLM-metric. For values close to 1 the histograms are sparsely filled. As a result, the extracted mapping becomes meaningless (indicated by the dashed line).

imates a certain part of an upright oriented object that is located somewhere in the scene together with its distance and height. Regions in the image containing no Stixels are implicitly understood as free space.

The Stixel computation is formulated as a MAP estimation problem, this way ensuring to obtain the best segmentation result for the given stereo input. An example result for this method is shown in Figure 1.

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

$$\begin{aligned} L &= \{L_u\}\,, \text{ with } 0 \leq u < w \\ L_u &= \{s_n\}\,, \text{ with } 1 \leq n \leq N_u \leq h \\ s_n &= \left\{v_n^b, v_n^t, c_n, f_n(v)\right\}, \text{ with } 0 \leq v_n^b \leq v_n^t < h\,,\, c_n \in C \end{aligned} \tag{5}$$

The total number of segments s_n 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. The term $f_n(v)$ is a function providing the depth of that segment at row position v (with $v_n^b \leq v \leq v_n^t$). All segments s_{n-1} and s_n are adjacent which implicitly guarantees that every image point is assigned to exactly one label.

Modeling all segments as piecewise planar surfaces simplifies the function set f_n to linear functions: object segments are assumed to have a constant disparity while ground segments follow the disparity gradient of the ground surface. This idea is illustrated in Figure 4.

Searching for the best Stixel segmentation L^* leads to a MAP estimation problem, such that:

$$L^* = \arg\max_{L \in \mathbb{L}} P\left(L \mid D\right). \tag{6}$$

Applying Bayes' theorem allows to rewrite the posterior $P\left(L \mid D\right)$ as $P\left(L \mid D\right) \sim P\left(D \mid L\right) \cdot P\left(L\right)$. The term $P\left(D \mid L\right)$ rates the probability of the input D given a labeling L and serves as the data term for the optimization. The second term $P\left(L\right)$ describes the overall probability for a possible labeling L and thus is the lever to incorporate prior world knowledge, *e.g.* ordering and gravity constraints. We discussed details on the latter term in [17].

Since this paper discusses how to efficiently take confidence cues into account, particular emphasis is put on the data term. $P\left(L \mid D\right)$ is written as the column-wise product

$$P\left(L \mid D\right) \sim \prod_{u=0}^{w-1} \underbrace{P\left(D_u \mid L_u\right)}_{\text{data term}} \cdot \underbrace{P\left(L_u\right)}_{\text{prior}}. \tag{7}$$

Next, the column-wise term $P\left(D_u \mid L_u\right)$ factorizes to

$$P\left(D_u \mid L_u\right) = \prod_{n=1}^{N_u} \prod_{v=v_n^b}^{v_n^t} \underbrace{P_{\mathrm{D}}\left(d_v \mid s_n, v\right)}_{\text{sensor model}}. \tag{8}$$

Here, $P_{\mathrm{D}}\left(d_v \mid s_n, v\right)$ represents the probability for a single disparity measurement d_v at image row coordinate v to belong to a possible Stixel segment s_n. To yield a robust estimation, P_{D} is defined as a mixture model composed of a Gaussian and a uniform distribution:

$$\begin{aligned} P_{\mathrm{D}}\left(d_v \mid s_n, v\right) = &\frac{p_{\text{out}}}{d_{\max} - d_{\min}} \\ &+ \frac{1 - p_{\text{out}}}{A_{\text{norm}}} e^{-\frac{1}{2}\left(\frac{d_v - f_n(v)}{\sigma^{c_n}(f_n, v)}\right)^2} \end{aligned} \tag{9}$$

By using the p_{out} parameter this forward sensor model [23] allows to consider that a disparity value might be an outlier. Originally, we modeled this eventuality as a global and thus measurement independent property. Although this seems applicable for most visual conditions, we know that in case of errors in the disparity estimation outliers do not just occur randomly. As shown in Figure 2, outliers often affect relatively large connected regions of the image. Accordingly, by not accurately modeling this characteristic, one runs the risk to have false object detections in these areas. Using the proposed method, the Stixel results for the scenarios shown in Figure 2 are given in Figure 5.

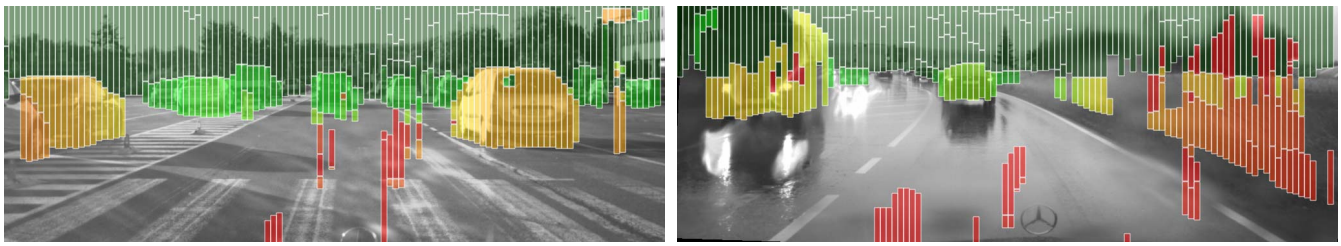

Figure 5: Stixel results for two visually challenging situations are using the approach of Pfeiffer *et al*. One can clearly see numerous false positives (mostly red) on the ground surface caused by stereo matching.

6. Using Confidences for Stixel Computation

For obtaining a more measurement specific outlier model, stereo confidence cues are used. As discussed in Section 4, these confidence cues are not used directly but are mapped to an outlier probability.

Once this step is carried out, the integration of the pixel-specific outlier probability into the original sensor model (*c.f*. Equation 9) is straightforward: Instead of processing the plain disparity measurement d_v in P_{D}, the tuple (d_v, p_v) is used which is the disparity measurement d_v along with the corresponding outlier probability p_v. The term p_{out} is replaced with p^*_{out} such that:

$$p^*_{\mathrm{out}} = p_v \left(1 - p^{\mathrm{min}}_{\mathrm{out}}\right) + p^{\mathrm{min}}_{\mathrm{out}} \tag{10}$$

As can be seen, the global outlier model parameter p_{out} is not just substituted by p_v. Instead, we ensure a lower bound, *i.e*. a minimum outlier probability $p^{\mathrm{min}}_{\mathrm{out}}$.

That decision is for two reasons: Firstly, when not providing stereo confidence cues, using $p_v = 0$ yields the original sensor model of Equation 9. Secondly and more crucial for our application, in awareness that the stereo confidence cue might not always be correct (*i.e*. $p_v \to 0$ even though d_v in fact is an outlier), this approach still enables to model a minimum outlier rate $p^{\mathrm{min}}_{\mathrm{out}}$.

This procedure is confirmed by our experiments. These show that modeling a rather defensive (and thus higher) outlier rate $p^{\mathrm{min}}_{\mathrm{out}}$ does not cause to miss objects in case the visual conditions are good. Thus, when striving for a system with maximum robustness, this certainly is a safe choice.

The results of using this extended sensor model for the scenarios in Figure 2 are given in Figure 6. The difference to the base line results obtained in Figure 5 is striking. No false positives are visible for the confidence version whereas many false positives occurred in the original version.

7. Using Confidences for Other Applications

Confidences are helpful for further applications driven by stereo vision. The following popular stereo-based tasks are easily extended to use confidence cues:

- Occupancy map generation: The disparities are triangulated and registered in a map. In addition to weighting a depth measurements with the sensor uncertainty [1], it can also be weighted with the inlier probability computed in Section 4.

- Ego-motion estimation / Registration: In extension to [15], a weighted ICP [18] can be used. The registration is performed using all triangulated disparities weighted with their inlier probabilities.

8. Evaluation on Real-World Traffic Data

Two types of data sets are used for evaluating the effectiveness of our approach. The first one is large: 308 sequences with more than $76,000$ frames in total. Those contain a mixture of $50\,\%$ normal daytime scenes and $50\,\%$ challenging scenes (rain, snow, night, reflections) thus overemphasizing sophisticated scenarios.

For this data set we rely on "weak" ground truth to measure the false positive rate: We assume that in all scenarios the driver maintains at least a one second time-gap to preceding vehicles or other objects. This should be true for all normal driving conditions. For all frames the driven corridor is projected one second ahead. All Stixels residing in that area are considered as false positives. In addition, RADAR data is used for reducing the driven corridor in those rare cases where the one second gap is under-run.

The second data set is much smaller and consists of 10 sequences with a total of $2,500$ frames. The particularity of this data set is that all 3D structures limiting the available free space are manually labeled thus allowing to obtain about $200,000$ ground truth Stixel measurements. This information allows to compute the detection rate which, of course, is an important aspect since minimizing the false positive rate can be achieved by simply not detecting any objects. Hence, the second data set helps to counterbalance this effect. For putting things into perspective, we also compute the false positive rate for the smaller data set. This ground truth data set is publicly available [1].

For the evaluation we work with three configurations:

[1] http://www.6d-vision.com/ground-truth-stixel-dataset

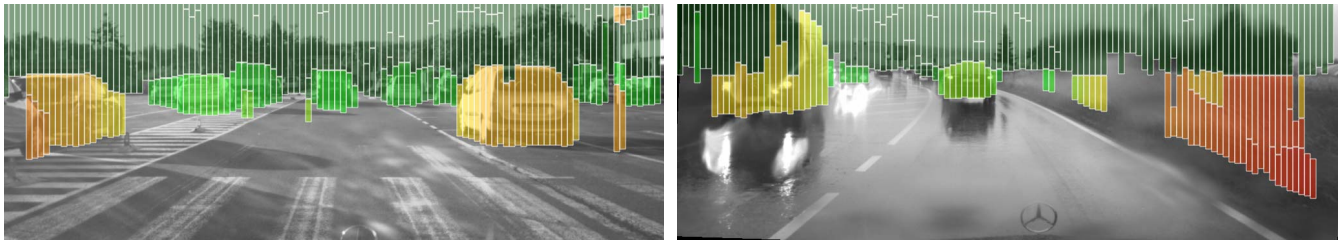

Figure 6: Results of our extensions to the Stixel computation of Pfeiffer *et al*. The difference to Figure 5 is striking. Using stereo confidence cues in the sensor model allows to remove nearly all false positives while the detection rate is kept high.

- Base line: SGM stereo and Stixels are computed according to [17]. No confidence information is used.
- Sparsification: SGM and the proposed confidence metrics are computed. A manually optimized confidence threshold is applied for discarding all depth measurements with a lower confidence from the disparity map. Then, this map is fed into the same Stixel engine.
- Stixels with confidences: SGM and the proposed confidence metrics are computed. Subsequently, we transfer both the disparity and the confidence map to the Stixel engine as described in Section 6.

The findings of these tests are summarized in Table 1. During evaluation the parameter configuration of the Stixel scheme was not altered. Any differences purely result from using the confidence metrics and the way how they are taken into account. The thresholds for sparsification have been chosen to optimize the trade-off between the false positive and detection rate which is $c_{min}^{LC} = 0.1$, $c_{min}^{PKRN} = 0.15$, and $c_{min}^{MLM} = 0.2$.

We obtain $2,055$ frames containing false positives in the base line approach. When using sparsification on the disparity input, the number of false positives is reduced to 637 with LC, 758 with PKRN, and 648 with MLM.

In comparison, the ground truth data set shows very similar results. The format is: frames with false positives (detection rate). The base line approach reaches $200\,(0.815)$, LC yields $73\,(0.801)$, PKRN has $83\,(0.797)$, and MLM reaches $79\,(0.804)$. Altogether, the three metrics enable to reduce the number of frames showing false positives roughly by a factor of three. LC and MLM operate on quite a similar performance level while PKRN is falling slightly behind.

The results when using confidence cues as suggested are as follows: On the large data set we obtain 360 frames with false positives when using LC, 719 with PKRN, and 301 in case of MLM. Respectively, for the ground truth data set the numbers are $45\,(0.802)$ for LC, $94\,(0.807)$ for PKRN, and $48\,(0.792)$ for MLM. Note that the detection rate improves by about $5\,\%$ for all experiments if frames showing windshield wipers are ignored. According to the findings of Section 4, the good performance of LC despite its simplicity is partly surprising. Our explanation is that most matching errors occur in low-textured areas where LC is most effective. Problems with repetitive structures, a known shortcoming of local stereo methods that is detected with PKRN and MLM, are rarely observed when using SGM.

In conclusion, exploiting stereo confidences throughout the whole processing chain clearly proves to have a positive effect. Compared to the base line, we yield a factor of up to six less false detections. Compared to sparsification the results still improve by a factor of two.

9. Conclusions and Outlook

In this contribution, we presented an improvement of the state-of-the-art 3D Stixel intermediate representation by exploiting stereo confidence information in a probabilistic fashion. It is shown that the intuitive approach to sparsify the disparity maps based on confidence allows to reduce the false positive rate by a factor of three. Instead of simply applying such a threshold, using confidences in a Bayesian manner yields an additional improvement by a factor of two while maintaining the same detection rate. These findings have been obtained from an extensive evaluation over a large data base containing more than $76,000$ frames of mostly challenging video material. A subset of this data base containing 3D ground truth object data is considered to be made publicly available.

The best performing metric "Local Curve", a quality measure for the sub-pixel curve fit, comes at no extra computational cost. The same holds true for integrating confidence information into the subsequent Stixel processing step. We are convinced that similar improvements can be achieved in other stereo-driven tasks.

Future work will further extend the usage of the aggregated confidences up to the object level. Also, when using Stixels with motion information, the identical concept can be applied for using optical flow confidence information.

References

[1] H. Badino, U. Franke, and R. Mester. Free space computation using stochastic occupancy grids and dynamic pro-

configuration	metric	number of fp	frames with fp	number of fp	frames with fp	detection rate
base line	none	7,905	2,055	754	200	**0.815**
sparsification	LC	1,293	637	169	73	0.801
	PKRN	3,170	758	277	83	0.797
	MLM	2,615	648	296	79	0.804
confidence mapping	LC	**703**	360	**107**	**45**	0.802
	PKRN	1,747	719	210	94	0.807
	MLM	727	**301**	113	48	0.792

(a) (b)

Table 1: For evaluating our extension of the Stixel computation scheme, we considered three different stereo confidence metrics and compared against both the base line approach of Pfeiffer *et al*. and the straight-forward way of using sparsification of the disparity map. Table a) shows the false positive (fp) results for the large data set (76,000 frames). Table b) shows the result for the smaller data set (2,500 frames) for which the detection rate is computed using ground truth object data. The following thresholds for sparsification were used: $c_{min}^{LC} = 0.1$, $c_{min}^{PKRN} = 0.15$, and $c_{min}^{MLM} = 0.2$.

gramming. In *Workshop on Dynamical Vision, ICCV*, Rio de Janeiro, Brazil, October 2007. 1, 2, 7

[2] H. Badino, U. Franke, and D. Pfeiffer. The Stixel World - A compact medium level representation of the 3D-world. In *DAGM*, pages 51–60, Jena, Germany, September 2009. 1, 2

[3] R. Benenson, R. Timofte, and L. Van Gool. Stixels estimation without depth map computation. In *IEEE CVVT:E2M at ICCV*, November 2011. 2

[4] G. Egnal, M. Mintz, and R. P. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image Vision Computing*, 22(12):943–957, 2004. 2

[5] A. E. Elfes and L. Matthies. Sensor integration for robot navigation: Combining sonar and stereo range data in a grid-based representation. In *28th IEEE Conference on Decision and Control*, 1987. 1, 2

[6] D. Gallup, M. Pollefeys, and J.-M. Frahm. 3d reconstruction using an n-layer heightmap. In *DAGM*, pages 1–10, Darmstadt, Germany, September 2010. 2

[7] S. Gehrig, F. Eberli, and T. Meyer. A real-time low-power stereo vision engine using semi-global matching. In *ICVS*, Liège, Belgium, October 2009. Springer-Verlag. 3

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE CVPR*, pages 3354–3361, Providence, USA, June 2012. 1

[9] R. Haeusler and R. Klette. Analysis of KITTI data for stereo analysis with stereo confidence measures. In *ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation*, pages 158–167, Florence, Italy, 2012. 2, 3

[10] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE CVPR*, pages 807–814, San Diego, USA, June 2005. 1, 3

[11] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *IEEE CVPR*, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society. 3

[12] X. Hu and P. Mordohai. Evaluation of stereo confidence indoors and outdoors. In *IEEE CVPR*, pages 1466–1473, San Francisco, CA, USA, June 2010. 1, 2, 3, 3

[13] D. Kondermann, S. Abraham, G. J. Brostow, W. Förstner, S. Gehrig, A. Imiya, B. Jähne, F. Klose, M. A. Magnor, H. Mayer, R. Mester, T. Pajdla, R. Reulke, and H. Zimmer. On performance analysis of optical flow algorithms. In *Theoretical Foundations of Computer Vision*, 2012. 4

[14] O. Mac Aodha, G. J. Brostow, and M. Pollefeys. Segmenting video into classes of algorithm-suitability. In *IEEE CVPR*, pages 1054–1061, San Francisco, CA, USA, June 2010. 4

[15] A. Milella and R. Siegwart. Stereo-based ego-motion estimation using pixel tracking and iterative closest point. In *ICVS*, pages 21–30, Bari, Italy, January 2006. 2, 7

[16] K. Mühlmann, D. Maier, J. Hesser, and R. Männer. Calculating dense disparity maps from color stereo images, an efficient implementation. *IJCV*, 47(1-3):79–88, 2002. 3

[17] D. Pfeiffer and U. Franke. Towards a global optimal multi-layer Stixel representation of dense 3D data. In *BMVC*, Dundee, Scotland, August 2011. 1, 2, 5, 5, 8

[18] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3D Digital Imaging and Modeling*, 2001. 7

[19] D. Scharstein and R. Szeliski. Middlebury online stereo evaluation, 2002. http://vision.middlebury.edu/stereo. 1, 4

[20] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 1, 3, 4

[21] F. Stein. The challenge of putting vision algorithms into a car. In *The Eighth IEEE Workshop on Embedded Vision, CVPR*, pages 89–94, Providence, RI, USA, June 2012. 5

[22] P. Steingrube, S. Gehrig, and U. Franke. Performance evaluation of stereo algorithms for automotive applications. In *ICVS*, pages 285–294, Liège, Belgium, October 2009. 2

[23] S. Thrun. Learning occupancy grid maps with forward sensor models. *Autonomous Robots*, 15:111–127, 2003. 2, 5

[24] A. Wedel, A. Meissner, C. Rabe, U. Franke, and D. Cremers. Detection and segmentation of independently moving objects from dense scene flow. In *EMMCVPR*, 2009. 3

[25] Y. Zhang, A. S. Dhua, S. J. Kiselewich, and W. A. Bauson. Challenges of embedded computer vision in automotive safety systems. In B. Kisačanin, S. S. Bhattacharyya, and S. Chai, editors, *Embedded Computer Vision*, Adv. in Pattern Recognition, pages 257–279. Springer London, 2009. 5

[26] Z. Zhang. A stereovision system for a planetary rover: Calibration, correlation, registration, and fusion. In *Proc. IEEE Workshop on Planetary Rover Technology and Systems*, Minneapolis, USA, April 1996. 2