Stereo Ground Truth With Error Bars

Daniel Kondermann[1], Rahul Nair[1], Stephan Meister[1], Wolfgang Mischler[1], Burkhard Güssefeld[1], Katrin Honauer[1], Sabine Hofmann[2], Claus Brenner[2] and Bernd Jähne[1]

1: Heidelberg Collaboratory for Image Processing at IWR,
Ruprecht-Karls-Universität Heidelberg
{firstname.lastname}@iwr.uni-heidelberg.de
2: Institute of Cartography and Geoinformatics, Leibniz Universität Hannover
{firstname.lastname}@ikg.uni-hannover.de

Abstract. Creating stereo ground truth based on real images is a measurement task. Measurements are never perfectly accurate: the depth at each pixel follows an error distribution. A common way to estimate the quality of measurements are error bars. In this paper we describe a methodology to add error bars to images of previously scanned static scenes. The main challenge for stereo ground truth error estimates based on such data is the nonlinear matching of 2D images to 3D points. Our method uses 2D feature quality, 3D point and calibration accuracy as well as covariance matrices of bundle adjustments. We sample the *reference data error* which is the 3D depth distribution of each point projected into 3D image space. The *disparity distribution* at each pixel location is then estimated by projecting samples of the reference data error on the 2D image plane. An analytical Gaussian error propagation is used to validate the results. As proof of concept, we created ground truth of an image sequence with 100 frames. Results show that disparity accuracies well below one pixel can be achieved, albeit with much large errors at depth discontinuities mainly caused by uncertain estimates of the camera location.

1 Introduction

Reference data is needed when quantitative performance evaluations are a requirement; this is for example the case for safety-relevant applications such as driver assistance systems. Whenever real data needs to be augmented with ground truth, measurement devices such as 3D scanners are used. These devices come with their own limits of accuracy. 3D scanners are for example limited in accuracy at objects with low reflectance, glossy surfaces or high geometric detail. Therefore, ground truth is never perfect - we need to understand the limits of the measurement devices in order to judge the quality of a ground truth dataset.

As a rule of thumb, a measurement device should be one order of magnitude more accurate than the required accuracy for the system to be evaluated. Many current stereo benchmarks analyze the number of pixels with a disparity error of one or more pixels [1]. Hence, to create stereo ground truth, the disparity map

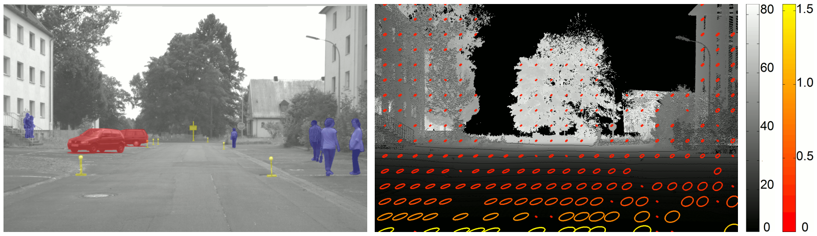

Fig. 1. Ground truth needs error bars. Left: left stereo image with overlay of dynamic objects. Right: ground truth disparities with sparse overlay of 3σ uncertainty ellipses. The disparity error is encoded in the color of the ellipse. Since the measured reference data is always subject to measurement errors the resulting ground truth dataset can contain uncertainties.

coming with the camera images should be around a tenth of a pixel accurate. The 3D data acquired by a scanner needs to be analyzed in pixel disparity space, resulting in mainly two errors: first, a depth-dependent error is introduced by the 3D-to-2D projection. This error becomes smaller with distance in case the scanner has a constant accuracy with respect to spatial coordinates. Second, a matching-dependent error occurs caused by bad alignment of the 2D image with the 3D scene. This results in very large, mostly bimodal error distribution near depth discontinuities. Both error sources cause highly different errors at individual pixel locations, rendering a general approximate error estimate for the full dataset relatively meaningless.

For this paper, we set up a high-end camera stereo system and reconstructed a large outdoor set using the best LIDAR system available for this task. Our aim is to focus on accuracy: how accurate can our real-world ground truth become at individual pixels when all involved systems are state of the art? To this end, we devised a method to estimate the accuracy of our ground truth at the pixel level. This paper does not propose a new dataset. Instead, we propose *a method* to create arbitrary stereo ground truth datasets with reliable per-pixel error bars (cf. Figure 1). Although our approach generalizes to arbitrary 3D scanners and camera setups in static scenes, we focus on large-scale outdoor scenes ($>$ $30.000\,\mathrm{m}^2$) which can to date only be acquired by LIDAR systems.

Our approach is illustrated in Figure 2 and is divided into the following steps: The static scene is scanned first and then a calibrated stereo sequence is recorded within this scene. The camera location for each frame is locally estimated based on manually selected 2D-3D-correspondences. All cameras and correspondences are inserted into a bundle adjustment model considering all error sources appropriately based on Gaussian errors in 2D feature localization, LIDAR accuracy and camera calibration parameters. Finally, the covariance of the functional is evaluated at the solution to assess the uncertainty in the derived camera extrinsics. The resulting error distributions of the inputs (LIDAR, image data, intrinsics) and derived inputs (extrinsics) are mapped into image space and, subsequently, into disparity space using both analytic error propagation and

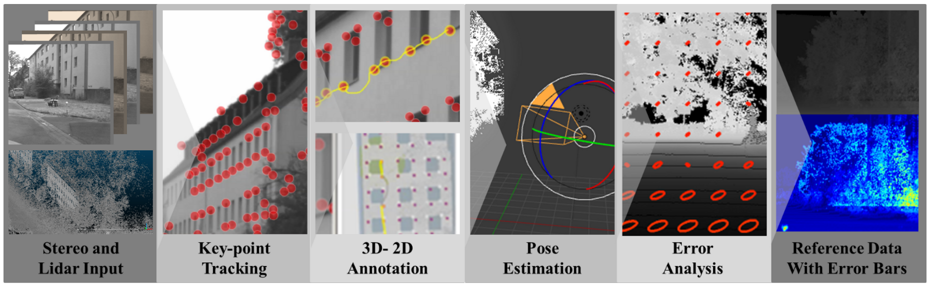

Fig. 2. Workflow stages: Starting with a LIDAR scan and an image sequence we compute 2D feature tracks. These are matched manually with landmark 3D points using manual annotations (Section 3.1). Using these annotations and the other 2D feature tracks we estimate the pose of each frame (Section 3.1). By means of covariance analysis and uncertainty propagation we then obtain uncertainties in the localization of reprojected 3D point cloud (Section 4). We then combine these localization uncertainties with the reprojections to finally output reference disparity maps and per pixel disparity distributions (Section 5).

Monte Carlo sampling. As a result, our method comprises a full error propagation, starting with Gaussian error assumptions of the involved measurement devices and ending at per-pixel non-parametric disparity distributions.

2 Related Work

Generation Techniques: Ground truth generation implies two parts: an evaluation dataset and a reference dataset with superior accuracy. Different techniques differ in the way these datasets are obtained [2].

Synthetic imagery [3–5] allows for generation of reference data with little uncertainty and makes white box testing of algorithms feasible by varying parameters such as geometry, light and materials. Yet, it remains to be shown whether content and renderer model reality well enough [6, 7].

Another option is to record real data and use *manual annotations*. While relatively new to low-level vision, efforts have been undertaken with some success [8]. With the advent of crowd-sourcing platforms [9], generation of such data has also become scalable. While the accuracy is reported to be good in general, possible biases introduced by humans are yet to be investigated.

Finally, reference data can also be obtained by *measurement* e.g by using more than two cameras [10] , additional devices such as the Kinect[11] a LIDAR scanner [12] or by using multiple exposures and UV-paint as in [1]. The approach of using more data sources is not as costly and sometimes scales very well because existing vision algorithms only need to be slightly modified. It should be noted however that in any case the reference data is itself obtained by measurement and therefore subject to uncertainty. Assessing this uncertainty in our opinion is of utmost importance as statements such as “LIDAR is always more accurate than stereo” do not hold in general [13].

Stereo Datasets[1]: General-purpose real-world reference data has been published in the Middlebury database [1] with an estimated accuracy of around 1/60th of a pixel. This value is derived from assumptions on the used block matching scheme and a down-sampling of originally larger images.

The EISATS database comprises a variety of sequences both real and synthetic [10, 14]. Using a third camera in the real dataset for additional redundancy proved to be beneficial for achieving an improved quality, but the accuracy of this data has not been thoroughly evaluated.

The closest approach to ours in terms of experimental setup is the KITTI dataset [12]: here, a stereo setup was combined with a car-mounted laser scanner. Mounting a LIDAR on the car has two main advantages. The scene can be recorded both in 2D and 3D at the same time and the density of 3D measurements is maximized as the LIDAR is very close to the optical axis of the stereo cameras. A disadvantage is that the system is moving while scanning, introducing a possibly low point density at high speed as well as motion artifacts. Although the accuracy was not explicitly evaluated in the original publication, it is reported by the authors to be less than three disparities for most of the pixels.

In our approach, the scene is scanned first and recorded later. Hence, motion artifacts cannot occur and the sampling is spatially roughly uniform. In both KITTI and our setup LIDAR was chosen as the most accurate and viable option to obtain depth in large scenes. Note, however, that our approach can be applied to any measurement technique with known uncertainty. Also all the main focus of all these databases is the creation of the ground truth database and the evaluation of algorithms. We focus on neither of both: Our aim is to exemplify error bar computation for real-world stereo ground truth using an appropriate statistical model.

Finally, the work most similar to ours in terms of scope is [13]. Here uncertainties in camera intrinsics/extrinsics, LIDAR measurements and image key-point estimation are propagated to obtain reconstruction uncertainties for multiple view stereo. While the authors make extensive use of sampling to estimate uncertainty we provide an analytical solution for both camera pose estimation and the uncertainty of the disparity maps. This makes handling large numbers of frames (more than 1000 vs 25 in [13]) tractable in the first place. While comparing a re-implemented version of their method with the proposed method we not only see a considerable speed up, even for small problems - we also observe tighter bounds on the camera pose uncertainty (cf. Section 3.2).

Uncertainty Estimation for Bundle Adjustment: A rich body of work exists on the theory of uncertainty estimation in the related field of bundle adjustment [15–18]. Most techniques use local features of the bundle adjustment energy in the optimum e.g. covariance analysis. A lot of effort is then put into

[1] Although most of the following works comprise additional datasets next to stereo data, we only focus on the latter.

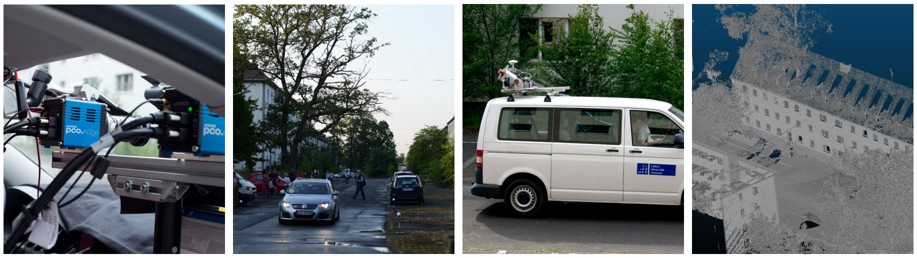

Fig. 3. From left to right: stereo rig, set photo, LIDAR mounted on car and resulting data.

tackling the inherent gauge ambiguity issue of the structure from motion problem. While we do use a bundle adjustment variant for estimating the camera parameters we circumvent the gauge ambiguity issue by fixing the gauge to the LIDAR reference frame. Also, it should be noted that our final goal is not the reconstruction of the camera parameters but rather stereo disparity maps with a per pixel uncertainty. To assess the quality of our camera reconstructions we build on work in [19].

3 Ground Truth Acquisition

The acquisition modalities are depicted in Figure 3. A reference 3D point cloud of a street of houses was collected using a RIEGL VMX-250-CS6. The stereo system consists of two cameras with a 30 cm baseline equipped with 12 mm lenses With a sensor size of 16.64 mm×14.04 mm, this corresponds to a field of view of 69.5°. The image sequences were acquired at 200 Hz with a resolution of 2560×1080 pixels. Preprocessing steps of the stereo data involved a lossless compression [20] of the 16 bit pixel data to 8 bits as well as camera calibration using [21]. Further details of the acquisition system can be found in the supplemental material.

3.1 2D-3D Alignment

All measurement based reference data acquisition systems rely on a 2D-3D alignment step at some point of the processing pipeline. To build on this step for both explaining our alignment process as well as on how we derive error bars, we will now review the basic pose estimation and calibration process.

With K we refer to the set of possible internal camera parameters and with $\mathbf{so}(3)$ the group of rotations. For a distortion free perspective camera with 4 parameters[2] $K = \mathbb{R}^4$. Let

$$\pi : (\mathbf{X}, \mathbf{t}, \kappa) \rightarrow \mathbf{x}, \tag{1}$$

$$\mathbf{X} \in \mathbb{R}^3, \mathbf{t} \in \mathbf{so}(3) \times \mathbb{R}^3, \kappa \in K, \tag{2}$$

[2] horizontal vertical focal lengths (f_x, f_y), principle point (c_x, c_y)

be the projective mapping of point $\mathbf{X}$ from the world to image coordinate system using the extrinsic parameters $\mathbf{t}$ and intrinsics κ. Furthermore, let $\{(\mathbf{X_i}, \mathbf{x_i^j})\}$ be a set of 3D-2D correspondences of p measured 3D points $\mathbf{X_i}$ and their projections $\mathbf{x_i^j}$ in the jth frame of an image sequence containing n images. Then, the optimal intrinsic parameter κ^* and set of extrinsics $T^* = \{\mathbf{t^j}^*\}$ for each of the n frames is given by

$$(T^*, \kappa^*) = \underset{T,\kappa}{\operatorname{argmin}} \sum_{j=0}^{n} \sum_{i \in V(j)} \left\| \pi\left(\mathbf{X_i}, \mathbf{t^j}, \kappa\right) - \mathbf{x_i^j} \right\|^2, \tag{3}$$

where $V(j) \subset [0...p]$ is the subset of 3D points that are visible in the jth frame. For a fixed camera - LIDAR setup such as KITTI this is done once in a calibration step with calibration targets before acquisition. Both geometry and projection of salient points are known here such that P can be obtained automatically. In our case the LIDAR and the camera rig measure independently. This has the advantage of having LIDAR data at a much higher point density and allows for capturing image sequences from other camera modalities (e.g Time-of-Flight, Plenoptic cameras) without requiring all cameras to be mounted on the same rig. In this setup, however, 2D-3D correspondences cannot be automatically aligned anymore. Picking individual points out of eight million options is an extremely tedious and error-prone task. We propose an annotation and processing pipeline minimizing the risk of false correspondences (cf. Figure 2).

2D-3D Correspondence estimation/annotation 2D feature tracks ($\mathbf{x_i^j}$) were automatically obtained with Voodoo Tracker[3] using the Harris Corner detector and a cross correlation based feature tracking.[4] A subset of the tracks was matched manually with 3D points. This is difficult since each point in the 2D projection of the cloud corresponds to many 3D points at different depths. One solution would be to automatically mesh the point clouds, but it turns out that current approaches do not work well enough on our kind of data and also modify the location of the points in a non-linear way introducing unknown biases to the measurements. To ease point picking, we reduced the 3D point cloud to a 2D representation in two steps:

Map Annotation We manually select landmark 3D points which were also always found by the 2D feature tracker. These points are visualized in a “fold-out” map of the measurement perimeter. Correspondences are established by connecting map landmarks with 2D features in the images.

Range Annotation Using an initial pose estimate computed from these correspondences, a range image with LIDAR reflectance information was created, containing at most one point per pixel from which additional correspondences can be chosen[5].

[3] http://www.digilab.uni-hannover.de/docs/manual.html

[4] Cross correlation window: 21×21. Search neighborhood 21×21.

[5] Screenshots and usage videos of the tools can be found in the supplemental material.

Camera Estimation With Known Variances Neither the feature tracks nor the 3D points or internal camera parameters are perfect. Also the intrinsic calibration routine usually delivers a good initial guess $\hat{\kappa}$ for the intrinsics. We assume Gaussian errors in each of these values:

$$\mathbf{X_i} = \mathbf{Z_i} + \epsilon_{\mathbf{X_i}} \ , \epsilon_{\mathbf{X_i}} \sim \mathcal{N}_3(0, \Sigma_{\mathbf{X_i}}) \quad (4)$$

$$\hat{\kappa} = \ \kappa + \epsilon_{\kappa} \ \ , \epsilon_{\kappa} \sim \mathcal{N}_4(0, \Sigma_{\kappa}) \quad (5)$$

$$\mathbf{x_i^j} = \ \mathbf{z_i^j} + \epsilon_{\mathbf{z_i^j}} \ , \epsilon_{\mathbf{x_i^j}} \sim \mathcal{N}_2(0, \Sigma_{\mathbf{x_i^j}}) \quad (6)$$

To accommodate for these errors we modify Equation 3:

$$(\{\mathbf{Z_i}\}^*, T^*, \kappa^*) = \underset{(\{\mathbf{Z_i}\}, T, \kappa)}{\operatorname{argmin}} \ \Phi(\{\mathbf{Z_i}\}, T, \kappa), \quad (7)$$

with

$$\begin{aligned} \Phi(\{\mathbf{Z_i}\}, T, \kappa) = \sum_{j=0}^{n} \sum_{i \in V(j)} \Bigg(& \left\| \pi\left(\mathbf{Z_i}, \mathbf{t^j}, \kappa\right) - \mathbf{x_i^j} \right\|^2_{\Sigma_{\mathbf{x_i^j}}} \\ & + \left\| \mathbf{X_i} - \mathbf{Z_i} \right\|^2_{\Sigma_{\mathbf{X_i}}} \\ & + \left\| \hat{\kappa} - \kappa \right\|^2_{\Sigma_{\kappa}} \Bigg). \end{aligned} \quad (8)$$

$\|.\|^2_{\Sigma}$ denotes the squared Mahalanobis distance. Note the quadratic penalty terms in Equation 8 and explicit usage of latent variables $\mathbf{Z_i}$ and κ. These are required as the first residual term is not linear in $\mathbf{X_i}$ and $\hat{\kappa}$ whereas it is in $\mathbf{x_i^j}$. This splitting of variables is often used to be able to better treat nonlinearities in Gaussian energy functionals. [22, 23]. Also, note that the first term corresponds to a bundle adjustment problem and the last two terms to priors on $\mathbf{X_i}$ and $\mathbf{x_i^j}$. In the optimization, it is therefore possible to include 2D feature tracks without 3D correspondences. Parameter estimation was done using the open source Ceres Solver [24] library.

3.2 Consistency And Precision of the Pose Estimation With Synthetic Data

To assess the precision and consistency of our pose estimation system we borrow ideas from [19]. Here, a method is proposed to compute consistency and precision of a dataset with respect to a reference dataset with lower but nonzero uncertainty. As the output of our system has the highest available precision we have to resort to synthetic data and make some changes to the formulas in [19] to cater to the zero uncertainty of our reference.

Consistency is a measure for the likelihood that both reference and synthetic datasets have the same parameters. As in [19] we report the Mahalanobis distance between the synthetic reference and the methods using the estimated pose covariance.

- **Precision** refers to the certainty of the method of the correctness of its parameter estimate. Given two parameter estimates with a similar consistency with regard to the reference, the estimate with the smaller uncertainty should be favoured. Here, we report the trace of the estimated covariances.

Table 1 summarizes the results. The reference data was generated by randomly picking p key points in the first frame, randomly choosing a depth for each key point between 5 and 70 meters and finally, by rejecting 3D points not visible in the $n-1$ other camera frames. The evaluation dataset was obtained by adding Gaussian noise according to the *noise* column on key point position and 3D point. We compare our method to a sampling based strategy similar to [13] where the 2D and 3D points are perturbed around the estimated solution (s times), the best new parameter set obtained by minimizing the bundle adjustment functional (keeping 2D and 3D measurements fixed) and estimating the sample mean and covariance. In the result columns, we report mean consistency, precision and run times in seconds after 30 runs. The standard deviation for consistency was always around 1, for precision and run time an order of magnitude smaller than the reported values. While we observe mostly similar consistency values between both methods - with the sampling consistency deteriorating with higher noise levels and larger datasets - our method produces a tighter precision bound on the parameter estimate with much faster run times. Further parameter sweeps can be found in the supplementary material.

4 Reference Data with Error Bars

Once the pose estimation in Equation 8 has been solved we can proceed in creating reference data by computing a range image based on κ, T and the LIDAR point cloud by means of Equation 1. This reference data contains holes with no information whenever no LIDAR measurements map to the corresponding pixel location. In the following we consider the extended reference data mapping

$$\tilde{\pi}_b : (\mathbf{X}, \mathbf{t}, \kappa) \rightarrow (\mathbf{x}, d) \tag{9}$$

which not only computes the projected image location of a 3D point but also the disparity of this point given stereo baseline b. With $\mathbf{d} = (\mathbf{x}, d)$ we will denote

Noise	Number of	Number of	Sampling	Sampling	
[cm, px]	points p	frames n	s = 100	s = 1000	Ours
(5, 0.1)	100	5	(5.1, 3.9e-4, 0.4)	(5.1, 4.0e-4, 4.7)	(5.3, 1.1e-4, 0.1)
(5, 0.5)	100	10	(7.7, 1.7e-2, 0.8)	(7.6, 1.7e-2, 8.2)	(7.6, 5.5e-3, 0.2)
(1, 0.1)	1000	10	(8.1, 7.8e-5, 9.5)	(7.9, 7.9e-5, 96)	(8.5, 2.2e-5, 2.4)
(5, 0.5)	1000	10	(8.2, 1.8e-3, 9.5)	(8.0, 1.8e-3, 97)	(7.2, 5.2e-4, 2.1)

Table 1. Pose estimation results on synthetic data. The tuples reported in the right 3 columns correspond to consistency, precision and run time. Lower values are better.

the vector containing image coordinates and disparity. We omit the subscript b in the further discussion as it remains constant for each sequence.

The inputs in $\tilde{\pi}(\mathbf{X_i}, \mathbf{t^j}, \kappa)$ are either measurements or values derived from measurements. As measurements always contain errors the reference point $\tilde{\pi}(...)$ will also have an error. To assess theses errors quantitatively we need to first obtain error estimates for $\mathbf{X_i}, \mathbf{t^j}$ and κ.

1. For the **3D point position $\mathbf{X_i}$** we assume that the components are independently distributed such that $\Sigma_{\mathbf{X_i}} = \sigma^2_{\mathbf{X_i}} I$. In our case this is the measurement error of the LIDAR scanner. For point clouds consisting of multiple LIDAR scans that were merged[12] via iterative closest points (ICP) or similar methods the error should be the error propagated from the ICP fit.
2. For the **camera pose $\mathbf{t^j}^*$** we assume that $\mathbf{t^j} \sim \mathcal{N}_6(\mathbf{t^j}^*, \Sigma_{\mathbf{t^j}})$. As $\mathbf{t^j}$ is a value derived from a least squares fit, $\Sigma_{\mathbf{t^j}}$ can be obtained by evaluating the covariance matrix of Φ at the solution $s^* = \{\mathbf{t^j}, ..., \}$ with
$$COV_{\Phi}(s^*) = (J_{\Phi(s*)} J^T_{\Phi(s*)})^{-1}. \tag{10}$$
Here, $J_{\Phi(s*)}$ is the Jacobian of the residual vector of Φ evaluated at solution s^*. $\Sigma_{\mathbf{t^j}}$ is the diagonal block of $COV_{\Phi}(s^*)$ corresponding to the parameter block belonging to $\mathbf{t^j}$. Note that a regular bundle adjustment scenario has an inherent scale ambiguity which leads to $J_{\Phi(s*)}$ being rank deficient. In contrast, our functional has full rank as the scale is given by the 2D - 3D correspondences. Also note that by supplying the correct error estimates during the alignment fit COV is properly scaled.
3. For the **camera intrinsics** κ we either use the same approach as chosen for $\mathbf{t^j}$ or use variances estimated by external calibration tools. Again the distribution is assumed to be Gaussian with $\kappa \sim \mathcal{N}_4(\kappa, \Sigma_\kappa)$.

The error distribution in $\tilde{\pi}$ of the reference point and the error in the disparity measure can be obtained via error propagation. This is achieved either via sampling input realizations from the above distributions or by analytical linear error propagation. For the latter, the full covariance matrix of the inputs evaluates to:

$$COV_{IN} = \begin{pmatrix} \Sigma_{\mathbf{X_i}} & & \\ & \Sigma_{\mathbf{t^j}} & \\ & & \Sigma_\kappa \end{pmatrix} \tag{11}$$

The error in $\tilde{\pi}$ is then obtained by linearizing $\tilde{\pi}$ at the reference point. Under assumption of a Gaussian distribution of the input variables the output is again Gaussian with covariance given by

$$COV_{\mathbf{d}} = J_{\tilde{\pi}(\mathbf{x},d))} COV_{IN} J^T_{\tilde{\pi}(\mathbf{x},d)}. \tag{12}$$

The choice between sampling and linear propagation depends on the available computational resources as sampling will deliver more accurate output error distributions given enough samples while linear error propagation is analytical and thus fast.

4.1 Reference Data Sensitivity

In the following we will give an analysis of our reference data using the tools provided above. We will first discuss the error values used for the inputs. The **LIDAR accuracy** is obtained from the data sheet as $\sigma_{\mathbf{X_i}} = 1\,\text{cm}$. We use this accuracy measure for the error propagation step. For the contribution of the 3D points towards pose uncertainty (cf. Eq. 10) we have to assume a larger error due to the point spacing. Therefore, the localization of a manually picked point (e.g. a window corner) is only accurate up to the mean distance between points. This was determined to be $\sigma'_{\mathbf{X_i}} = 3.5\,\text{cm}$ by estimating the point density on building facades where the landmark points were chosen from. The **feature track accuracy** was empirically estimated to be $\sigma_{xij} = 0.5\,\text{px}$, while errors in **focal length and principal point** were obtained from our calibration routine as $\sigma_{\kappa(f_x,f_y)} = 1.97\,\text{px}$ for the focal length and $\sigma_{\kappa(c_x,c_y)} = 1.46\,\text{px}$ for the principal point. For the **pose estimation accuracy**, we report the square root of the diagonal entries of $\Sigma_{\mathbf{t_j}}$ obtained from covariance analysis to be $(r_x, r_y, r_z) = (3, 3, 2)\times 10^{-4}$ for the rotation and $(t_x, t_y, t_z) = (1.23, 2.53, 2.17)\,\text{cm}$ for the translation over 100 frames. The rotation is parametrized using a three dimensional angle-axis representation. The error has an upper bound[6] of 0.026°. For a LIDAR point at 50 m distance this corresponds to a localization error of around 2 cm. The error in the translation also amounts to 2 cm. Using the errors obtained from the input we can compute the uncertainty in the reference data by means of error propagation. For each reference point the full covariance in $\mathbf{d}$ (i.e. pixel localization and disparity error) was computed using both linear error propagation and sampling. In Figure 4 the square roots of the diagonal entries are reported for an example scene. The first two rows correspond to the localization error and the third row is the disparity error in logarithmic scale. For both linear propagation and sampling we see the expected inverse distance reduction of all errors. While the disparity error for most parts is under a pixel the localization error exceeds five pixels for points closer than a few meters. Also noticeable is the rise in x localization error towards the image edges observable in all our sequences. We believe that this is related to a rotational error of the camera localization. Finally, we can see by comparing sampling and linear propagation that the sampling propagation in general gives a tighter bound on the reference data error while preserving the general shape. As both propagation methods yield similar results we conclude that linear error propagation can be used to obtain a quick though looser bound on the reference data error.

5 Disparity Maps with Error Bars

So far we discussed the reference data quality in terms of the localization and disparity error of each reference point. For evaluating a stereo algorithm we are faced with a slightly different question as we are concerned with the question how good a given disparity map is. We hence need a distribution of possible disparity

[6] Based on the maximum deviation of the angle-axis vector

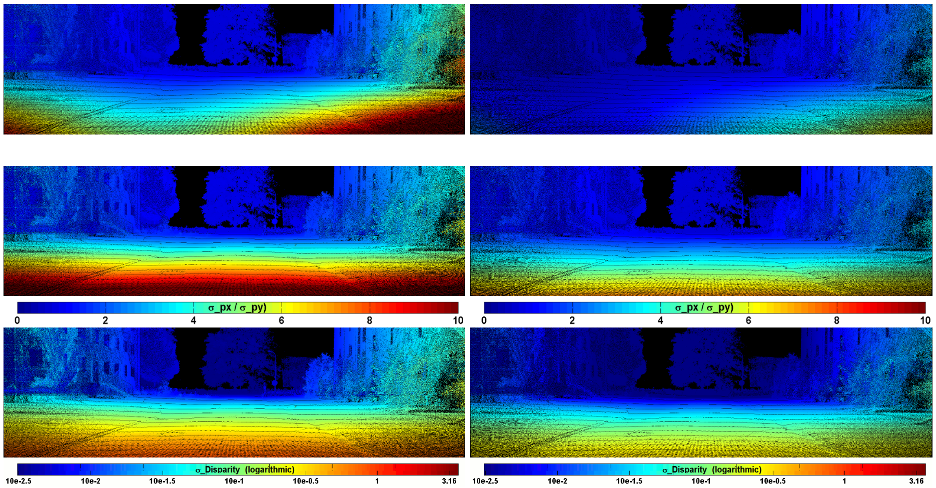

Fig. 4. Diagonal entries of uncertainty $\Sigma_{\mathbf{d}}$ obtained by **linear error propagation (left)** and **sampling (right)**. From top to bottom: Localization error in x and y as well as disparity error of reference data points. Note that the bottom row is scaled logarithmically. While the general form of the error distribution is the same for both analytic and sampling based propagation, we obtain tighter bounds on all errors using sampling.

values in each pixel. Given a set of reference data points with uncertainty $R = \{(\mu_{\mathbf{r}}, \Sigma_r)\}$ computed as described in Section 4, we define the probability of a disparity map $\mathbf{D}$ to be

$$p(\mathbf{D}|R) = \prod_{\mathbf{x_i} \in \mathbf{D}} \frac{1}{N} \sum_{(\mu_{\mathbf{r}}, \Sigma_r) \in R} \exp\left((\mathbf{x_i} - \mu_{\mathbf{r}})^T \Sigma_r^{-1} (\mathbf{x_i} - \mu_{\mathbf{r}})\right) \tag{13}$$

with $\mathbf{x_i} = (\mathbf{p_i}, d)$ the disparity d at pixel position $\mathbf{p_i}$ and normalization N. The Gaussian distribution in Eq. 13 is multivariate (in pixel position and disparity). This distribution can alternatively be computed by either sampling from the reference data distribution or analytically from the input data distribution directly using Gaussian error propagation. The main drawback of a linear error propagation is that the projection of Gaussian disparity distribution into image space yields multi-modal per-pixel distributions which cannot be accounted for using linear propagation. Figure 5 shows such distributions at example pixel locations. We can distinguish three error cases: first, due to extrinsic camera parameter uncertainty the locations of depth edges are projected to different pixel locations. This causes bimodal disparity distributions since either the background or the foreground is sampled. The result is a very high variance, i.e. a large, though correct error bar on the ground truth.

Second, multi-modal distributions can occur caused by back surfaces: multiple surfaces such as the front and back of a house as well as the houses in

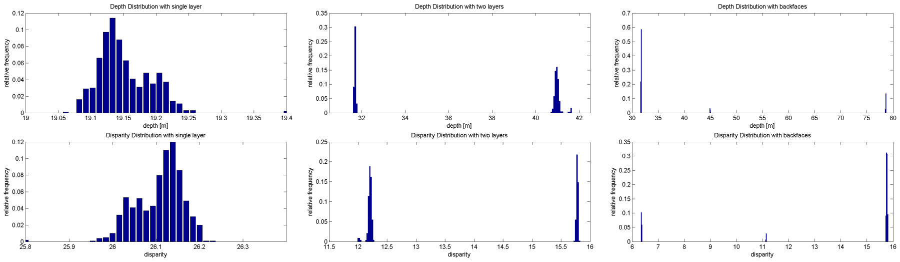

Fig. 5. Example distributions on sampled depth maps (1000 samples). From left to right: pixel with single depth layer, edge pixel with two depth layers, pixel with unresolved back faces. Top row: depth distribution. Bottom row: disparity distribution.

the background of the LIDAR point cloud are projected to the same pixel. This is a fundamental limitation of point clouds - yet established meshing tools can not deal with our data as was explained in Section 3.1. In these situations, the ground truth is not wrong per se - but more reasoning is required to decide whether the multi-modality of the distribution is caused either by a depth edge or back surfaces.

Third, in case the scanner did not measure a foreground object, for example due to limited resolution (landlines, small twigs on trees), the disparity distribution becomes unimodal but still displays the wrong depth of the object behind the small foreground object. This case can only be dealt with by more accurate measurement devices which not yet exist at least for our application. The problem can only be alleviated by manual segmentation of foreground objects which are visible in the image, but not in the 3D scan.

Once the per-pixel distributions in disparity space are sampled, we can reduce their information to per-pixel scalar values. Figure 6 displays two such options: the top image contains the median of the disparity distribution. Assuming that the number of foreground samples outweighs the number of back-surfaces by a factor of at least two, this is a robust ground truth depth. Note however, that this approach fails at depth boundaries when foreground and background can easily become equally likely. Therefore, the lower image displays the standard deviation of the disparity distributions. We can for example use it to define a ground truth mask as is common for stereo benchmarks such as Middlebury or KITTI: we choose a threshold defining when we cannot trust the ground truth any longer. To obtain meaningful ground truth for a pixel-accurate algorithm, one would typically choose a maximum standard deviation of 1 pixels.

It is important to mention that this type of masking is not necessarily the best option for performance evaluation. A simple performance metric based on the full distribution could be $m(\mathbf{D}|R) = -\log(p(\mathbf{D}|R))$. For reference data with localization error much smaller than the pixel size the sum in Equation 13 can be replaced with a single normal distribution belonging to the reference point in the respective pixel. The negative logarithm of the term then yields a per-pixel weighted sum of a squared distance metric. A more appropriate evaluation would

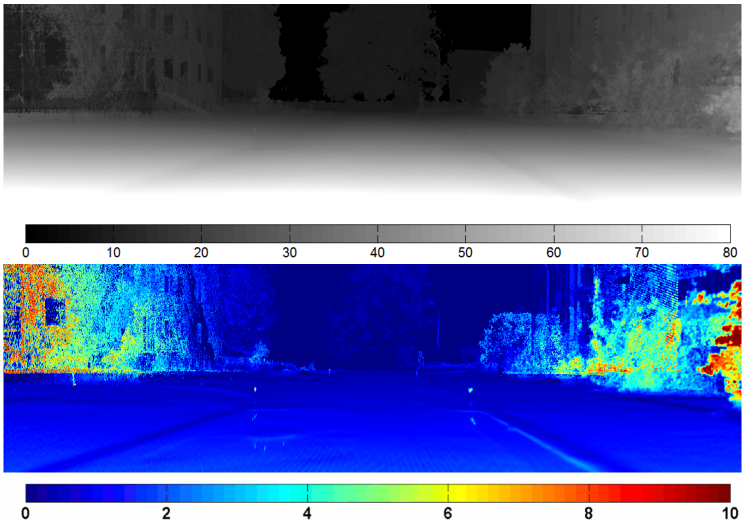

Fig. 6. Top: median of disparity distribution. Bottom: standard deviation of disparity distribution. High variances show regions with unreliable ground truth mainly caused by vegetation and camera misalignments. Regions looking like artifacts are caused by backsurfaces as explained in the text. In all other regions, the standard deviation is below two disparities.

require the stereo algorithm to propose a disparity distribution as well; then, the performance metric would compare ground truth and computed disparity distribution e.g. by a Kolmogorov-Smirnov test.

6 Conclusion and Outlook

We have presented a methodology to add error bars to image sequences with disparity ground truth. It is based on previously measured point clouds and arbitrary calibrated cameras and therefore highly versatile for all kinds of indoor as well as outdoor applications. However, due to the chosen 3D scanning device our approach is limited to static scenes.

Based on intuitive inputs such as calibration, 2D feature and 3D LIDAR accuracy we estimated the covariance matrix of our model at the solution to derive per-pixel depth-distributions. The results were used to define error bars, e.g. by computing the depth variance at each pixel.

Results with a recently recorded scene showed that the localization error caused by suboptimal camera estimates significantly deteriorates quality by introducing multi-modal depth distributions at depth edges, especially at objects close to the camera. Even with arguably the best hardware available today and highly tuned manual alignment tools, the disparity standard deviation exceeds several pixels at nearby objects. Objects with high geometric detail cannot be measured with LIDAR reliably, causing additional artifacts in the ground truth. In this paper we used the accuracy claimed in the LIDAR manufacturer's data sheet which should be a very good approximation. More detailed studies such as [25] will be incorporated in future work. Only in the background accuracies well below one pixel can be achieved. This indicates that a per-pixel quality estimate of real-world ground truth is very important for ground truth generation and any subsequent performance evaluation. Especially algorithms claiming to be pixel-accurate should only take into account a masked subset of the ground truth with standard deviations of less than 1 pixels. It should be noted however that thresholding the reference data is only one simple way of harnessing known error distributions of reference data for purposes of performance analysis. By analyzing not only the absolute difference between stereo output $\mathbf{D}_s$ with reference depth image $\mathbf{D}_r$

$$\mathbf{R} = |\mathbf{D_R} - \mathbf{D_S}|, \tag{14}$$

but also taking into account a consistency value inspired by the Mahalanobis distance used in Section 3.2

$$\mathbf{C} = |\mathbf{D_R} - \mathbf{D_S}|/\mathbf{S_R}, \tag{15}$$

where $\mathbf{S_R}$ is the interquartile range of the reference data distribution, it is possible to gain more insights into the performance characteristics of a stereo algorithm; especially, it is possible to identify situations where the algorithm is achieving the same accuracy as the reference data measurements yet other areas where no statements can be made about the algorithm performance. We give a more detailed discussion of the metrics as an outlook in the supplementary material, as the results presented there are only intended as a proof of concept and require further investigation to be conclusive.

In terms of our experimental setup, the accuracy could be improved in smaller scenes by using our approach with a micrometer-accurate structured light scanner delivering object meshes rather than point clouds. Then, the limiting factor becomes camera pose estimation, which is a matter of future studies. We will further add ground truth with error bars for optical flow and look at improved methods for backface analysis of large point clouds such as manual meshing, usage of camera motion and point normal analysis. The results and presented here as well as a supplementary video are available on the dataset homepage[7].

Acknowledgments We thank Wolfgang Niehsen and his Team at Robert Bosch GmbH, Computer Vision Research Lab, Hildesheim, for supplying the test car, camera mount and tons of input regarding meaningful content of the scenes we recorded. We further thank Jens Taupadel, Jakob Knauer and Moritz Wandsleb at Universität Hannover for acquiring and processing the scans. Finally, we thank our lab members Karsten Krispin, Alexandro Sanchez-Bach, Ekaterina Melnik for their assistance in data processing, Florian Becker and Frank Lenzen for helpful discussions as well as AEON Verlag & Studio GmbH for the organization of all helpers and facilities.

[7] http://hci.iwr.uni-heidelberg.de/Benchmarks/document/StereoErrorBars/

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision **92** (2011) 1–31
2. Kondermann, D.: Ground truth design principles: an overview. In: Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications, ACM (2013) 5
3. Onkarappa, N., Sappa, A.D.: Synthetic sequences and ground-truth flow field generation for algorithm validation. Multimedia Tools and Applications (2013) 1–15
4. Haltakov, V., Unger, C., Ilic, S.: Framework for generation of synthetic ground truth data for driver assistance applications. In: GCPR. (2013)
5. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), ed.: European Conf. on Computer Vision (ECCV). Part IV, LNCS 7577, Springer-Verlag (2012) 611–625
6. Meister, S., Kondermann, D.: Real versus realistically rendered scenes for optical flow evaluation. In: Proceedings of 14th ITG Conference on Electronic Media Technology, Informatik Centrum Dortmund e.V. (2011)
7. Güssefeld, B., Kondermann, D., Schwartz, C., Klein, R.: Are reflectance field renderings appropriate for optical flow evaluation? In: IEEE International Conference on Image Processing 2014 (ICIP 2014), Paris, France (2014)
8. Liu, C., Freeman, W.T., Adelson, E.H., Weiss, Y.: Human-assisted motion annotation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR08) **0** (2008) 1–8
9. Donath, A., Kondermann, D.: Is crowdsourcing for optical flow ground truth generation feasible? Proc. International Conference on Vision Systems (2013)
10. Morales, S., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. In: Computer Analysis of Images and Patterns, Springer (2009) 1078–1086
11. Meister, S., Izadi, S., Kohli, P., Hämmerle, M., Rother, C., Kondermann, D.: When can we use kinectfusion for ground truth acquisition? Proc. Workshop on Color-Depth Camera Fusion in Robotics (2012)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR), Providence, USA (2012)
13. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE (2008) 1–8
14. Vaudrey, T., Rabe, C., Klette, R., Milburn, J.: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In: Proc. of 23rd International on Conference Image and Vision Computing New Zealand, (IVCNZ08). (2008) 1–6
15. Kanatani, K.: Statistical optimization for geometric fitting: Theoretical accuracy bound and high order error analysis. International Journal of Computer Vision **80** (2008) 167–188
16. Kanatani, K.: Uncertainty modeling and model selection for geometric inference. Pattern Analysis and Machine Intelligence, IEEE Transactions on **26** (2004) 1307–1319

17. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment, a modern synthesis. In: Vision algorithms: theory and practice. Springer (2000) 298–372
18. Förstner, W.: Reliability analysis of parameter estimation in linear models with applications to mensuration problems in computer vision. Computer Vision, Graphics, and Image Processing **40** (1987) 273–310
19. Dickscheid, T., Läbe, T., Förstner, W.: Benchmarking automatic bundle adjustment results. In: 21st Congress of the International Society for Photogrammetry and Remote Sensing (ISPRS). (2008) 7–12, Part B3a
20. Jähne, B.: Digitale Bildverarbeitung. 7 edn. Springer, Berlin (2012)
21. Abraham, S., Hau, T.: Towards autonomous high-precision calibration of digital cameras. In: Videometrics V, Proceedings of SPIE Annual Meeting. Volume 3174., Citeseer (1997) 82–93
22. Afonso, M.V., Bioucas-Dias, J.M., Figueiredo, M.A.: Fast image recovery using variable splitting and constrained optimization. Image Processing, IEEE Transactions on **19** (2010) 2345–2356
23. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv- l 1 optical flow. In: Pattern Recognition, (Proc. DAGM07). Volume 4713 of LNCS. (2007) 214–223
24. Agarwal, S., Mierle, K., Others: Ceres solver. (http://ceres-solver.org)
25. Boehler, W., Bordas Vicent, M., Marbs, A.: Investigating laser scanner accuracy. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences **34** (2003) 696–701