MULTI-VIEW STEREO AND LIDAR FOR OUTDOOR SCENE MODELLING

C. Strecha[a], W. von Hansen[b], L. Van Gool[a,c] and U. Thoennessen[b]

[a] KU-Leuven ESAT/PSI, Belgium, {christoph.strecha,luc.vangool}@esat.kuleuven.be
[b] FGAN-FOM Ettlingen, Germany {wvhansen,thoe}@fom.fgan.de
[c] ETH Zürich BIWI, Switzerland

KEY WORDS: LIDAR , self-calibration, evaluation, multi-view stereo, Bayesian generative model based stereo

ABSTRACT:

In this paper we want to start the discussion on whether image based 3-D modelling techniques and especially multi-view stereo can possibly be used to replace LIDAR systems for outdoor 3D data acquisition. Two main issues have to be addressed in this context: (i) camera self-calibration and (ii) dense multi-view depth estimation. To investigate both, we have acquired test data from outdoor scenes with LIDAR and cameras. Using the LIDAR data as reference we provide an evaluation procedure to these two major parts of the 3D model building pipeline. The test images are available for the community as benchmark data.

1 INTRODUCTION

Several techniques to measure the shape of objects in 3-D are available. The most common systems are based on active stereo, passive stereo, time of flight laser measurements (LIDAR) or NMR imaging. For measurements in laboratories, one usually puts small objects *into* a relatively large measurement device. In contrast to this, outdoor measurements require to bring the measurement equipment *to* considerably larger objects. Traditionally, close range photogrammetry had been used to capture objects of architectural or archaeological interest such as buildings. This technique required precisely calibrated cameras as well as a lot of manual work on expensive coordinate measurement devices to produce a result. A second technique for precise outdoor measurements that has become important during the past years is the application of LIDAR systems. In contrast to image based techniques, these are able to directly produce a 3-D point cloud based on distance measurements with an accuracy of less than 1 cm. The downside are high costs for the system and – compared to taking images with a camera – data acquisition is time consuming.

Automatic reconstruction from multiple view imagery already is a low-cost alternative to traditional methods, but could even become a replacement to LIDAR systems once the geometrical accuracy of the results could be proven. For this purpose we have acquired LIDAR data and high resolution images for several outdoor scenes. The LIDAR data will serve as geometrical ground truth to evaluate the quality of the image based reconstructions.

To tackle the problem of 3-D data acquisition from images, one has to consider camera self-calibration techniques and multi-view stereo. Camera calibration includes the computation of external (position and orientation) and internal camera parameters. Multi-view stereo takes this information to compute a dense 3-D model of the scene. Both problems are usually seen separate such that we opted to provide evaluation data and an evaluation procedure to measure their accuracy independently. We consider this as an first step to the necessary evaluation which considers both problems in common[1]. As an example we provide evaluation results for camera self-calibration (EPOCH, 2007) and a multi-view stereo approach that is based on generative models (Strecha et al., 2004, Strecha et al., 2006). The latter consists of two stages.

[1]Working systems that automatically produce 3-D models from uncalibrated images are rarely available yet.

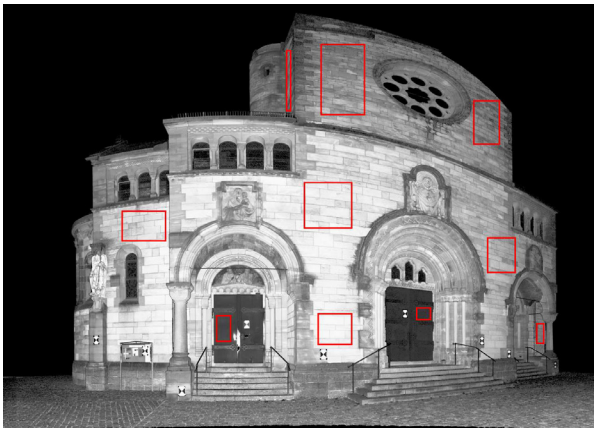

Figure 1: Example of the LIDAR data: intensity image of the church with marked planar regions.

Firstly, a global formulation consideres all possible depth and visibility configurations of the scene (Strecha et al., 2006). This approach is relatively memory and time consuming and is therefore suited only for small resolution images. Its advantage is the ability to find a good, often global, solution without depth initialisation (Strecha, 2007). In a second stage, the solution of this global formulation is used as an initialisation to a PDE-based approach that is applied to large scale images (Strecha et al., 2004).

With this contribution we want to start the discussion on whether multi-view stereo can compete with LIDAR in terms of accuracy. Note, that we do not treat the LIDAR data as an *absolute* ground truth reference. The LIDAR itself contains errors, that are estimated and translated to the image based acquisition. As a result we evaluate image based depth modelling relative to the accuracy of the LIDAR data. This is different from previous ground truth data sets, where the reference data is assumed to be given without errors (Scharstein and Szeliski, 2002, Seitz et al., 2006, Bellmann et al., 2007). A second difference is the choice for outdoor scenes. Whereas indoor scenes can be measured by active stereo systems, *e.g.* by using structured light, this is not an option for outdoor environments. The evaluation of passive, image based 3-D modelling techniques is in that sense especially for outdoor scenes of large practical importance.

The paper is organised as follows. Sec. 2 deals with the LIDAR data aquisition, *i.e.* the registration of different scans and the ac-

Ambiguity interval		53.5 m
Range resolution		0.8 mm
Range noise	10 m 20%	3.0 mm rms
@ distance,	10 m 100%	1.3 mm rms
reflectivity	25 m 20%	9.0 mm rms
	25 m 100%	3.0 mm rms
Field of view (v×h)		$310^\circ \times 360^\circ$
Angular resolution		$0.018^\circ \times 0.01^\circ$
Angular accuracy		$0.02^\circ \times 0.02^\circ$ rms
Beam divergence		0.22 mrad
Points per dataset		200 million

Table 1: Specifications of the IMAGER 5003.

curacy estimation. We link the LIDAR point clouds to the images by estimating the camera parameters and depth maps together with all variances in sec. 3. In sec. 4 we consider the evaluation procedure and show example results. Sec. 5 concludes the paper.

2 ACQUISITION OF GROUND TRUTH DATA

2.1 Data acquisition and registration

The datasource for ground truth in our project is terrestrial laser scanning (LIDAR). A laser beam is scanned across the object surface, measuring the distance to the object for each position. Modern systems also record the returned intensity of the laser beam so that a gray scale texture image is available. We had a Zoller+Fröhlich IMAGER 5003 laser scanner at our disposition (tab. 1). Multiple scan positions are required for complex object surfaces to handle missing parts due to occlusions. Even though fully automatic methods exist for registration (Akca, 2003, Bae, 2006, Rabbani and van den Heuvel, 2005, von Hansen, 2006), we have chosen a semi-automatic way utilising software provided by the manufacturer in order to get a clean reference.

2-D targets are put into the scene and marked interactively in the datasets. Then the centre coordinates are automatically computed and a registration module computes a least squares estimation of the parameters for a rigid transform between the datasets. Poorly defined targets can be detected and manually removed. The resulting mean accuracy for a single target was 1.1 mm for the church and 1.5 mm for the fountain dataset. The targets are visible in the camera images as well and are used to link LIDAR and camera coordinate systems.

2.2 Preparation of point clouds

First, some filters from the software are applied to mask out bad points resulting from measurements into the sky, mixed pixels and other error sources. Some regions containing clutter have been manually masked out and a 3×3 median filter had been applied to the data to remove remaining noisy points.

Then, a triangle mesh has been created separately from each point cloud so that image rays can be intersected with object surfaces. The scan data is acquired in a system of lines and rows so that direct neighbors are known. Four points on a square can generate zero to two triangles. The decision is based on the angle α between the viewing ray and the normal vector of the triangle. In order to avoid the inclusion of depth edges as surfaces, triangles are only generated when $\cos\alpha > 0.5$. In the ambiguous case of two triangles, the configuration that minimises the larger of the two angles is chosen.

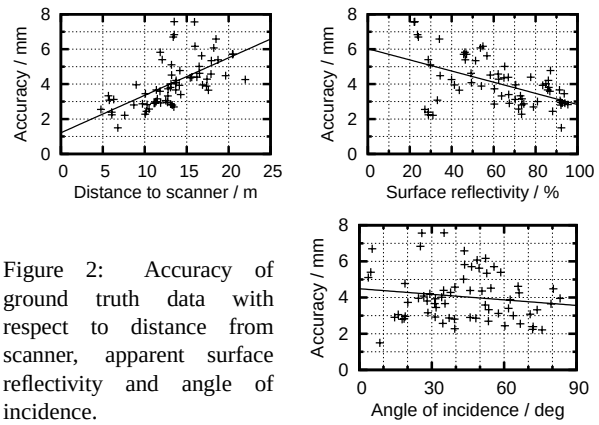

Figure 2: Accuracy of ground truth data with respect to distance from scanner, apparent surface reflectivity and angle of incidence.

2.3 Evaluation of point cloud accuracy

The manufacturer of the LIDAR system gives an accuracy of 1.3–9.0 mm for the range measurement based on distance and reflectivity (tab. 1). In this section, we will investigate the accuracy of our datasets. From scene knowledge, a total of 66 planar surface patches has been manually marked (fig. 1). For all the 3-D points inside each selection, a principal component analysis is performed. The covariance matrix of an ideal plane has rank two, for a real plane the smallest eigenvalue equals to the variance perpendicular to the plane. The associated eigenvector is the surface normal. As further parameters, the mean point of the surface as distance to the scanner and the mean gray value as estimate for the surface reflectivity are computed. Results are shown in fig. 2 separately for each parameter and are ordered by the amount of influence on the accuracy. The relationship between parameter and accuracy has been estimated by a linear function.

The most important factor is the distance from the scanner to the surface. For this reason, a maximum distance of 20–25 m has been used as cut-off during data export. The linear approximation ranges from 2–6 mm over the used range interval, yielding a variation of 4 mm. The apparent surface reflectivity is derived from the amount of light received by the sensor. These levels do not have a real physical meaning at this stage, but nevertheless covered the range from 20% to 95% without any adjustments. As with the distance, a relationship between reflectivity and accuracy exists, but the magnitude is slightly smaller. The linear approximation ranges from 3.0–5.5 mm, yielding only 2.5 mm variation. The last parameter is the angle of the surface normal with respect to the laser beam. An angle of 0° denotes a surface perpendicular to the light ray. This parameter is uncorrelated to the surface accuracy. One should be aware that the parameters are actually correlated, *e.g.* the apparent reflectivity depends on both distance and angle of incidence.

For all 66 regions chosen, the accuracy ranges from 1.5 mm for the best case to 7.6 mm for the worst case. The worst accuracies occur for regions on the church doors, which have a dark and reflective metal surface. Our results are in accordance to the accuracies given in tab. 1. From the diagrams an average accuracy of about 4 mm can be derived. Even though this value has been computed from planar regions, it can be generalised to any smooth object surface. In comparison to the accuracy of the targets (≤ 1.5 mm), we see that the registration error is smaller than the surface accuracy. It can therefore be concluded, that the registration is good enough as to not disturb the results.

The accuracy of 4 mm given here for the test data is an absolute accuracy. This is a high quality result as can be seen from the relative accuracy: Object size and typical measurement distances

Figure 3: Example of target measurements with their (enlarged) covariance for the church data set.

are in the range of 15 m, leading to a relative accuracy of $3 \cdot 10^{-4}$. For comparison, the object size of Seitz *et al*. (Seitz et al., 2006) has been 0.1 m and the accuracy given was 0.25 mm, leading to a relative accuracy that is about ten times worse.

3 GROUND TRUTH ESTIMATION FROM LIDAR

Together with the LIDAR data the scenes have been captured with a Canon D60 digital camera with a resolution of 3072×2048 square pixels. In this section we describe how the images are calibrated and how the ground truth 3-D model is obtained by using the LIDAR data. Our focus is thereby not only on the ground truth estimation itself but also on the accuracy of our ground truth data. The LIDAR depth estimates are themselves the result of an erroneous measurement process and as such only given in a certain error range. We propagate these errors into our image based ground truth estimation.

Errors for the multi-view stereo evaluation are introduced by: *(i)* the 3-D accuracy of the LIDAR data itself and *(ii)* by the calibration errors of the input cameras. The latter does influence the quality of multi-view stereo reconstructions strongly. Evaluations that take these calibration errors into account should therefore be based on per image reference depth maps (more details are given in sec. 4.2) as opposed to Seitz *et al*. (Seitz et al., 2006), who evaluate the stereo reconstructions by the Euclidean 3-D distance between estimated and ground truth triangle mesh.

3.1 Mean and variance of the camera calibration

LIDAR data and camera images are linked via targets that are visible in both datasets. Thus the laser scanner provides 3-D reference coordinates that can be used to compute the calibration parameters for each camera. For the camera calibration we assume a perspective camera model with radial distortion (Hartley and Zisserman, 2000). The images are taken without changing the focal length, such that the internal camera parameters $\boldsymbol{\theta}_{int} = \{f, s, x_0, a, y_0, k_1, k_2\}$ ($\mathbf{K}$-matrix and radial distortion parameters $k_{1,2}$) are assumed to be constant for all images. The external camera parameters are the position and orientation of the camera described by 6 parameters $\boldsymbol{\theta}_{ext} = \{\alpha, \beta, \gamma, t_x, t_y, t_z\}$. The total number of parameters $\boldsymbol{\theta}$ for N images is thus $7+6N$. To calibrate the cameras we used M targets, which have been placed in the scene. The 3-D position $\boldsymbol{Y}_j; j = 1 \ldots M$ and the covariance $\boldsymbol{\Sigma}_Y$ for these is provided by the laser scan software. In each input image i we estimated the 2-D position $\boldsymbol{y}_{ij}$ of these targets as it is shown in fig. (3) and used unit covariance $\boldsymbol{\Sigma}_{ij}$.

Let $\mathbf{y}$ denote all measurements, *i.e*. the collection of 3-D points $\mathbf{Y}_j$ and the 2-D image measurements $\mathbf{y}_{ij}$. The expected value of the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{int}, \boldsymbol{\theta}_{ext_1}, \ldots, \boldsymbol{\theta}_{ext_N}\}$ can be written as:

$$\boldsymbol{E}[\boldsymbol{\theta}] = \int p(\mathbf{y}') p(\boldsymbol{\theta}' \,|\, \mathbf{y}') \; \boldsymbol{\theta}' \; d\mathbf{y}' \; d\boldsymbol{\theta}'. \tag{1}$$

Here $p(\mathbf{y}')$ is the likelihood of data, *i.e*. among all 3-D points $\mathbf{Y}'_i$ and image measurements $\mathbf{y}'_{ij}$ only those will have a large likelihood that are close to the estimated values $\mathbf{y}$:

$$\begin{aligned} p(\mathbf{y}'_{ij}) &\propto \exp\left(-0.5(\mathbf{y}_{ij} - \mathbf{y}'_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1} (\mathbf{y}_{ij} - \mathbf{y}'_{ij})\right) \\ p(\mathbf{Y}'_j) &\propto \exp\left(-0.5(\mathbf{Y}_j - \mathbf{Y}'_j)^T \boldsymbol{\Sigma}_Y^{-1} (\mathbf{Y}_j - \mathbf{Y}'_j)\right) \end{aligned} \tag{2}$$

The second term, $p(\boldsymbol{\theta} \mid \mathbf{y}')$, is the likelihood of the calibration. This distribution is Gaussian and reflects the accuracy of the calibration, given the data points $\mathbf{y}'$. This accuracy is based on the reprojection error:

$$e(\boldsymbol{\theta}) = \sum_i^N \sum_j^M \left(\boldsymbol{P_i}(\boldsymbol{\theta}) \boldsymbol{Y}_j - \mathbf{y}_{ij}\right)^T \boldsymbol{\Sigma}_{ij}^{-1} \left(\boldsymbol{P_i}(\boldsymbol{\theta}) \boldsymbol{Y}_j - \mathbf{y}_{ij}\right) \;,$$

where $\boldsymbol{P_i}(\boldsymbol{\theta})$ projects a 3-D point $\boldsymbol{Y}_j$ to the image pixel $\mathbf{y}'_{ij}$ and the calibration likelihood becomes:

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto \exp\left(-0.5 e(\boldsymbol{\theta})\right) \;. \tag{3}$$

The covariance $\boldsymbol{\Sigma}$ of the camera parameters is similarly given by:

$$\boldsymbol{\Sigma} = \int p(\mathbf{y}') p(\boldsymbol{\theta}' \,|\, \mathbf{y}') \; \left(\boldsymbol{E}(\boldsymbol{\theta}') - \boldsymbol{\theta}'\right) \left(\boldsymbol{E}(\boldsymbol{\theta}') - \boldsymbol{\theta}'\right)^T \; d\mathbf{y}' \; d\boldsymbol{\theta}' \;. \tag{4}$$

To compute the solution of eqn. (1) and (4), we apply a sampling strategy. The measurement distribution $p(\mathbf{y})$ is randomly sampled. Given a specific sample $\mathbf{y}'$ the parameters $\boldsymbol{\theta}'$ are computed by a bundle adjustmentas the ML estimate of eq. (3):

$$\boldsymbol{\theta}' = \arg\max_{\boldsymbol{\theta}} \{\log p(\boldsymbol{\theta} \,|\, \mathbf{y}')\} \;. \tag{5}$$

Using eq. (5) and eq. (2) we can approximate the expected values and the covariance in eqn. (1) and (4) by a weighted sum over the sample estimates. As a result we obtain all camera parameters $\boldsymbol{\theta}$ by $\boldsymbol{E}[\boldsymbol{\theta}]$ and their covariance $\boldsymbol{\Sigma}$.

3.2 Mean and variance of ground truth depth

Given the mean and variance of the camera calibration we are now in the position to estimate the expected value of the per pixel depth and variance. Again we sample the camera parameter distribution given by $\mathbf{E}[\boldsymbol{\theta}]$ and $\boldsymbol{\Sigma}$ in eq. (1) and eq. (4):

$$p(\boldsymbol{\theta}') = \frac{\exp\left(-\frac{1}{2} \left(\mathbf{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}'\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}'\right)\right)}{2\pi^{\frac{7+6N}{2}} \mid \boldsymbol{\Sigma} \mid} \;, \tag{6}$$

and collect sufficient statistics for the per pixel depth values by the first intersection of the laser scan triangle mesh with the camera ray through each pixel. The result is the mean $\mathcal{D}^{ij}_{\text{LIDAR}}$ and variance $\mathcal{D}^{ij}_{\sigma}$ of the depth value for all pixels i in all cameras j. An example is shown in fig. 4 for one camera. Note, that this procedure allows to evaluate multi-view stereo reconstructions independent on the accuracy of the camera calibration. If the performance of the stereo algorithm is evaluated in 3-D (*e.g*. by the Euclidean distance to the ground truth triangle mesh (Seitz et al., 2006)) the accuracy of the camera calibration and the accuracy of the stereo algorithm is mixed. Here, the evaluation is relative to calibration accuracy, *i.e*. pixels with a large depth variance, given

Figure 4: Mean $\mathcal{D}_{\text{LIDAR}}$ (top) and variance $\mathcal{D}_\sigma$ (bottom) of the depth with respect to one camera. Red pixels denote pixels for which the camera ray does not intersect the laser scan model.

the uncertainty of the calibration, will influence the evaluation criterion accordingly. The variance plot in fig. 4 shows that large depth variance pixels appear near depth boundaries and for surface parts with a large slant. Obviously, these depth values vary most with a varying camera position. The reference depth maps (mean) and their variance will be used for evaluation in sec. 4.2.

4 DEPTH MODELLING FROM IMAGES

The 3-D modelling from high resolution images as the only input has made a huge step forward in being accurate and applicable to real scenes. Various authors propose a so called structure and motion pipeline (Akbarzadeh et al., 2006, Nister, 2004, Pollefeys et al., 2004, Rodriguez et al., 2005, Snavely et al., 2006, Strecha et al., 2003). This pipeline consists of mainly three steps. In the first step, the raw images undergo a sparse-feature based matching procedure. Matching is often based on invariant feature detectors (Mikolajczyk et al., 2005) and descriptors (Mikolajczyk and Schmid, 2005) which are applied to pairs of input images. Secondly, the position and orientation as well as the internal camera parameters are obtained by self-calibration (Hartley and Zisserman, 2000). The third step takes the input images, which have often been corrected for radial distortion, and the camera parameters and establishes dense correspondences or the complete 3-D model (see (Seitz et al., 2006) for an overview). In the remainder of this section we discuss and evaluate self-calibration and multi-view stereo on these ground truth data (fig. 3 and fig. 5).

4.1 Camera calibration

To compare results of self-calibration techniques based on our ground truth data we first have to align both camera tracks by a

Figure 5: Example image from the fountain data set.

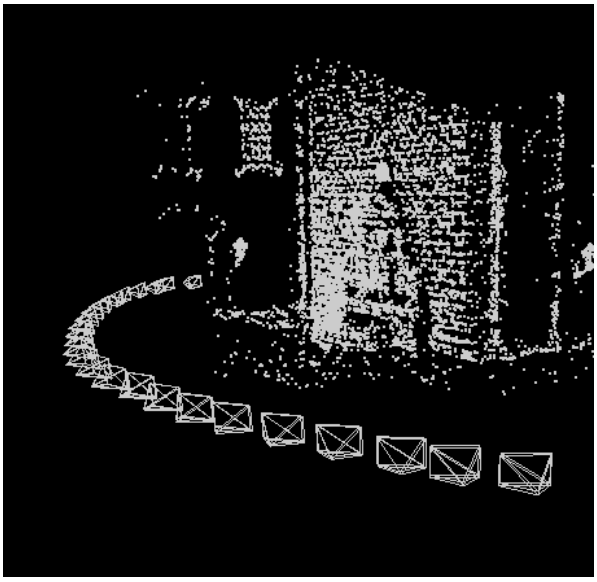

Figure 6: Camera calibration for the fountain data set in fig. 5

rigid 3-D transformation (scale, rotation and translation). First the scale is estimated by using the largest camera distance within each camera track. For one camera we estimate rotation and translation. These values are used as an initialisation to a non-linear optimisation that brings the evaluation track into the coordinate system of the ground truth track. After alignment the error of a certain calibration $\boldsymbol{\theta}_{eval}$ is given by the negative log-likelihood of eq. (6):

$$\epsilon = (\mathbf{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}_{eval})^T \boldsymbol{\Sigma}^{-1} (\mathbf{E}[\boldsymbol{\theta}] - \boldsymbol{\theta}_{eval}) \; . \qquad (7)$$

As an example we tested the calibration with the EPOCH (EPOCH, 2007) software which successfully calibrated all cameras for the fountain data set. The result of this automatic camera calibration is shown in fig. 6, where we plot the position and orientation of the cameras (both ground truth and the EPOCH cameras). Fig. 7 shows the difference in (x,y,z) position for each of the 25 cameras w.r.t. the ground truth position. The error bars indicate the 3σ value of our ground truth camera positions. Note, the larger variance for the first camera positions in this data set. They are due to a larger distance to the scene which results in less reliable camera positions. The camera positions lie within one cm from the ground truth and the average distance is 5 times larger than the position variance of the ground truth.

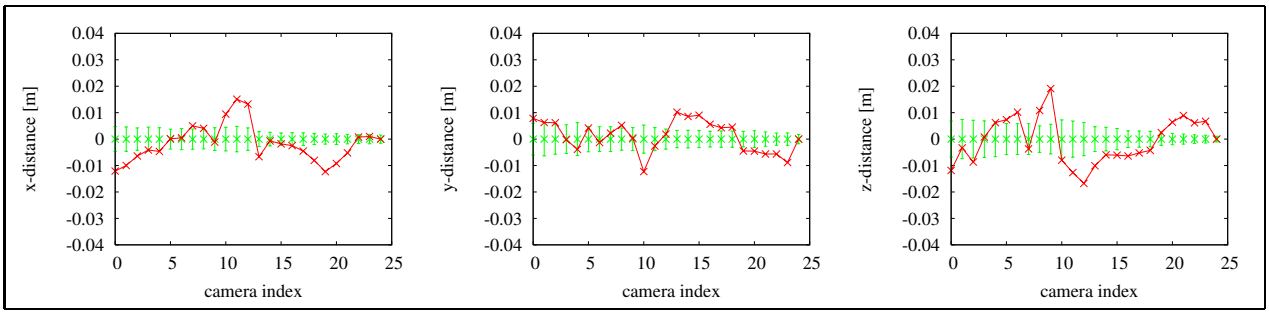

Figure 7: Position error [m] of the self-calibration (red) for the fountain data set in fig. 5. The green error bars indicate the 3σ value of the ground truth positions.

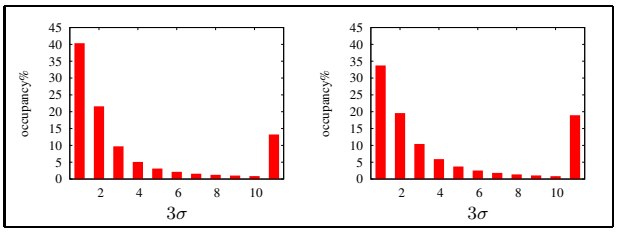

Figure 8: Histogram of the depth errors for the fountain (left) and church data set (right).

4.2 Multi-view stereo

Dense multi-view stereo applied to outdoor scenes cannot rely on visual hulls that are very useful for indoor stereo applications as evaluated by Seitz *et al*. (Seitz et al., 2006). Our test images have high resolution in order to meet the LIDAR precision and do not capture the object from all around. During data acquisition we also encountered problems due to pedestrians and changing light conditions. These aspects form a particular challenge for outdoor stereo reconstructions.

As input to the multi-view stereo algorithms we provide the ground truth camera calibration as estimated in sec. 3, the images, which have been corrected for radial distortion, the bounding volume of the scene as well as the minimal/maximal depth value w.r.t. to each input image. We evaluate the results for our generative model based multi-view stereo approach as formulated in (Strecha et al., 2004, Strecha et al., 2006, Strecha, 2007).

The results of this multi-view stereo approach are shown in fig. 9 for the fountain data. The accuracy of the stereo reconstruction $\mathcal{D}^{ij}_{Stereo}$ is evaluated by building a histogram $\boldsymbol{h}_k$ over the relative errors:

$$\boldsymbol{h}_k \propto \sum_{ij} \delta_k \left(\mid \mathcal{D}^{ij}_{\text{LIDAR}} - \mathcal{D}^{ij}_{Stereo} \mid, \mathcal{D}^{ij}_{\sigma} \right) . \qquad (8)$$

$\mathcal{D}^{ij}_{\text{LIDAR}}$ is the expected value of the LIDAR depth estimate at pixel position i and camera j and $\mathcal{D}^{ij}_{\sigma}$ its corresponding variance (both shown in fig. 4). Furthermore, $\delta_k()$ is an indicator function which evaluates to 1 if the depth difference $\mid \mathcal{D}^{ij}_{\text{LIDAR}} - \mathcal{D}^{ij}_{Stereo} \mid$ falls within the variance range $[k\mathcal{D}^{ij}_{\sigma}, (k+1)\mathcal{D}^{ij}_{\sigma}]$ and evaluates to zero otherwise. The stereo estimate $\mathcal{D}^{ij}_{Stereo}$ is obtained from a 3-D triangle mesh by computing the depth of the first triangle intersection with the j^{th} camera ray going through pixel i. Its value is given directly for multi-view stereo formulations that use a depth-map representation (Seitz et al., 2006). All depth estimates for which the absolute difference with the ground truth is larger than $30\mathcal{D}^{ij}_{\sigma}$ *and* all pixels for which the multi-view stereo reconstruction does not give depth estimates are collected all together in the last bin. These are all pixels indicated by blue in fig. 9. The relative error histogram for the fountain and church data are show in fig. 8. They can be interpreted as follows: $\sim 40\%$ ($\sim 34\%$) of the stereo depth estimates lie within the 3σ range of the LIDAR data for the fountain (church) data set; for $\sim 15\%$ ($\sim 20\%$) either no stereo depth exists or the error is larger than 30 times the LIDAR variance. Note, that our evaluation is relative to the camera calibration accuracy. This is achieved by evaluating the depth accuracy in the image by a local, pixel dependent value of $\mathcal{D}^{ij}_{\sigma}$. Untextured renderings of the multi-view stereo and LIDAR models are shown in fig. 10.

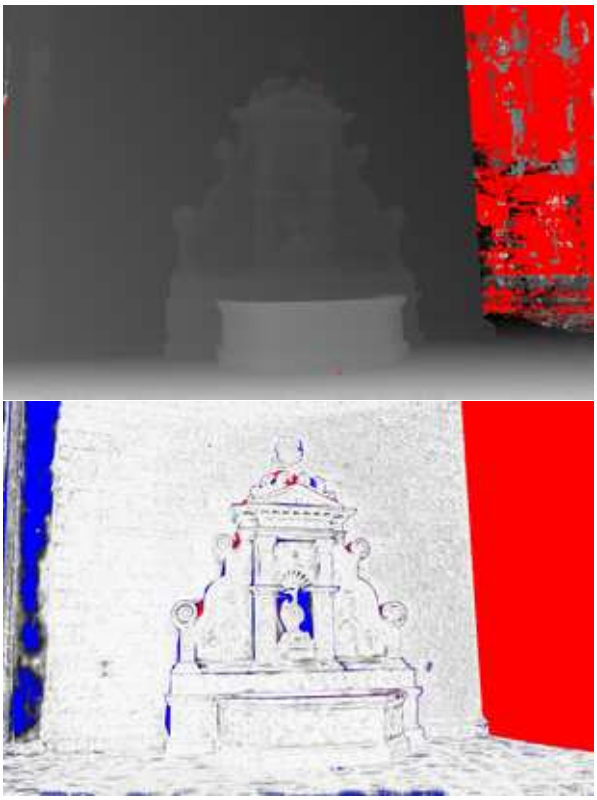

Figure 9: Stereo depth-map (top) and the variance weighted depth difference w.r.t. to the LIDAR ground truth (bottom): white pixels indicate small relative error; missing ground truth depth values are in red; blue pixels denote errors for which $\mid \mathcal{D}_{\text{LIDAR}} - \mathcal{D}_{Stereo} \mid / \mathcal{D}_{\sigma} \geq 30$.

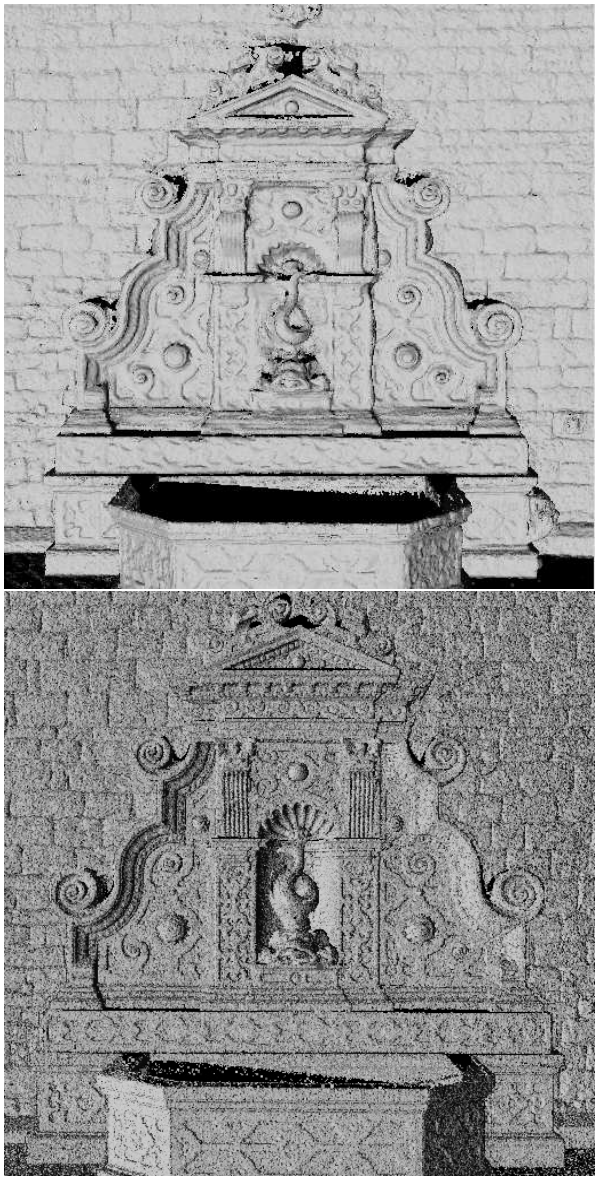

Figure 10: Multi-view stereo (top) and LIDAR (bottom) renderings of the fountain scene.

5 SUMMARY AND CONCLUSIONS

In this paper we investigated the possibility to evaluate 3-D modelling of outdoor scenes based on high resolution digital images using ground truth acquired by a LIDAR system. An evaluation carried out on the LIDAR data showed an average geometric accuracy of the surface measurements of about 4 mm and less than 1.5 mm. This is slightly better than the accuracy achieved from the images so that we can therefore conclude that our ground truth is of sufficient quality for further tests.

The link between LIDAR point cloud and images has been established via targets visible in both datasets and by generating per pixel depth maps with associated accuracies for each of the images. This allows to take into account the individual camera configuration so that poorly defined 3-D points will not deteriorate the evaluation result.

Example results have been shown for the the two important parts of image based 3-D acquisition, *i.e.* camera self-calibration and multi-view stereo. The reference data has been computed such that each of these steps can be evaluated separately and relative to the accuracy of the reference data. We will make the images available as benchmark data in order to help to enhance the accuracy of automatic image based 3-D modelling techniques. We hope that vision systems become available that can compete with LIDAR systems in terms of geometric quality.

REFERENCES

Akbarzadeh, A., Frahm, J., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Yang, R., Welch, G., Towles, H., Nister, D. and Pollefeys, M., 2006. Towards urban 3D reconstruction from video. In: *Proc. 3DPVT* , pp. 1–8.

Akca, D., 2003. Full automatic registration of laser scanner point clouds. In: Optical 3-D Measurements VI, pp. 330–337.

Bae, K.-H., 2006. Automated Registration of Unorganised Point Clouds from Terrestrial Laser Scanners. PhD thesis, Curtin University of Technology.

Bellmann, A., Hellwich, O., Rodehorst, V. and Yilmaz, U., 2007. A benchmarking dataset for performance evaluation of automatic surface reconstruction algorithms. In: ISPRS Workshop BenCOS.

EPOCH, 2007.
www.esat.kuleuven.be/˜visit3d/webservice/html.

Hartley, R. and Zisserman, A., 2000. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521623049.

Mikolajczyk, K. and Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE Transactions on PAMI* 27(10), pp. 1615–1630.

Mikolajczyk, K., Tuytelaars.T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and Van Gool, L., 2005. A comparison of affine region detectors. *IJCV* 65(1-2), pp. 43–72.

Nister, D., 2004. Automatic passive recovery of 3D from images and video. In: *Proc. 3DPVT* , IEEE Computer Society, Washington, DC, USA, pp. 438–445.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J. and Koch, R., 2004. Visual modeling with a hand-held camera. *IJCV* 59(3), pp. 207–232.

Rabbani, T. and van den Heuvel, F., 2005. Automatic point cloud registration using constrained search for corresponding objects. In: 7th Conference on Optical 3-D Measurements.

Rodriguez, T., Sturm, P., Gargallo, P., Guilbert, N., Heyden, A., Menendez, J. and Ronda, J., 2005. Photorealistic 3D reconstruction from handheld cameras. Machine Vision and Applications 16(4), pp. 246–257.

Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV* 47(1/2/3), pp. 7–42.

Seitz, S., Curless, B., Diebel, J., Scharstein, D. and Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: *Proc. CVPR* , IEEE Computer Society, Washington, DC, USA, pp. 519–528.

Snavely, N., Seitz, S. and Szeliski, R., 2006. Photo tourism: exploring photo collections in 3D. In: SIGGRAPH '06, ACM Press, New York, NY, USA, pp. 835–846.

Strecha, C., 2007. Multi-view stereo as an inverse inverence problem. PhD thesis, PSI-Visics, KU-Leuven.

Strecha, C., Fransens, R. and Van Gool, L., 2004. Wide-baseline stereo from multiple views: a probabilistic account. *Proc. CVPR* 1, pp. 552–559.

Strecha, C., Fransens, R. and Van Gool, L., 2006. Combined depth and outlier estimation in multi-view stereo. *Proc. CVPR* pp. 2394–2401.

Strecha, C., Tuytelaars, T. and Van Gool, L., 2003. Dense matching of multiple wide-baseline views. In: *Proc. ICCV* , pp. 1194–1201.

von Hansen, W., 2006. Robust automatic marker-free registration of terrestrial scan data. In: W. Förstner and R. Steffen (eds), Photogrammetric Computer Vision, IAPRS, Vol. XXXVI Part 3.