Evaluation of Distance-Aware KinFu Algorithm for Stereo Outdoor Data

Hani Javan Hemmat, Egor Bondarev, Gijs Dubbelman and Peter H. N. de With
Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands

Keywords: 3D Reconstruction, Stereo Camera Depth Data, Voxel-Models, Camera-pose Estimation, Weighting Strategy, Truncated Signed Distance Function (TSDF).

Abstract: In this paper, we report on experiments on deployment of an extended distance-aware KinFu algorithm, designed to generate 3D model from Kinect data, onto depth frames extracted from stereo camera data. The proposed idea allows to suppress the Kinect usage limitation for outdoor sensing due to the IR interference with sunlight. Besides this, exploiting the stereo data enables a hybrid 3D reconstruction system capable of switching between the Kinect depth frames and stereo data depending on the quality and quantity of the 3D and visual features on a scene. While the nature of the stereo sensing and the Kinect depth sensing is completely different, the stereo camera and the Kinect show similar sensitivity to distance capturing. We have evaluated the stereo-based 3D reconstruction with the extended KinFu algorithm with the following distance aware weighting strategies: (a) *weight definition* to prioritize importance of the sensed data depending on its accuracy, and (b) *model updating* to decide about the level of influence of the new data on the existing 3D model. The qualitative comparison of the resulting outdoor 3D models shows higher accuracy and smoothness of models obtained by introduced distance-aware strategies. The quantitative analysis reveals that applying the proposed weighting strategies onto stereo datasets enables to increase robustness of the pose-estimation algorithm and its endurance by factor of two.

1 INTRODUCTION

Accurate 3D reconstruction and mapping has been addressed as a vital topic and is playing a prominent role in such important research domains as 3D shape acquisition and modelling, surface generation and texturing, localization and robot vision (Engelhard et al., 2011; Newcombe et al., 2011a; Steinbrucker et al., 2011; Whelan et al., 2012a; Whelan et al., 2013). During recent years, the advent of powerful general-purpose GPUs has resulted in the first generation of real-time 3D-reconstruction applications which use depth data obtained from a low-cost depth Kinect sensor (PrimeSense; Kinect; Asus) to generate 3D geometry for relatively large and complex indoor environments (Newcombe et al., 2011b; Izadi et al., 2011; Bondarev et al., 2013; Whelan et al., 2012b; Whelan et al., 2012a; Whelan et al., 2013). Since the depth sensor is based on projected Infra-Red (IR) patterns, it is almost impossible to sense outdoor scenes during daylight. This can be explained by the IR interference of the sensor and the sunlight. While the Kinect is unable to sense the outdoor environments, a stereo camera can provide depth data for an outdoor scene and can be potentially used by the KinFu algorithm to reconstruct 3D model of outdoor environment.

The most prominent real-time 3D reconstruction applications, such as KinectFusion (Newcombe et al., 2011b; Izadi et al., 2011), Kintinious (Whelan et al., 2012b), the open source KinFu (PCL, 2011) and KinFu Large Scale (Bondarev et al., 2013), utilize the low-cost depth sensor (Newcombe et al., 2011b; Izadi et al., 2011) to sense the environment and reconstruct the corresponding 3D model based on the TSDF voxel-model. While recent evaluation of the Kinect intrinsics has revealed relative robustness to ambient light, incidence angle, and radiometric influences, the important sensor limitation is the low accuracy for large distance measurements (Chow et al., 2012; Khoshelham, 2011; Khoshelham and Elberink, 2012). Since the default weighting strategy of the TSDF model is not capable to preserve more accurate data against less accurate data, an improved approach has been proposed by introduction of *weight definitions* and *updating strategies* (Javan Hemmat et al., 2014a). The resulting improvements in model quality (Javan Hemmat et al., 2014a) and pose-estimation accuracy (Javan Hemmat et al., 2014b) obtained by these strategies, have motivated us to extend the ap-

Javan Hemmat H., Bondarev E., Dubbelman G. and de With P..
Evaluation of Distance-Aware KinFu Algorithm for Stereo Outdoor Data.
DOI: 10.5220/0004901507460751
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (PANORAMA-2014), pages 746-751
ISBN: 978-989-758-004-8

proach to the domain of stereo sensor data.

In this paper, we extend the capability of KinFu to fuse data from stereo sensors, which enables 3D model reconstruction of outdoor scenes. We evaluate the impact of the distance-aware weighting strategies on the quality of the resulting KinFu 3D model obtained from the stereo-based depth data. For this, we generate depth data from the input stereo camera data and feed the depth data to the extended distance-aware KinFu algorithm. Finally, we perform quantitative and qualitative comparison of the obtained 3D models against models generated by the conventional KinFu algorithm.

The paper is structured as follows. Section 2 describes the extended weighting strategies. Section 3 elaborates on performed experiments. Section 4 provides analysis and discussion of the results. Section 5 concludes the paper.

2 WEIGHTING STRATEGIES

2.1 Conventional TSDF Model

In the original TSDF model (Curless and Levoy, 1996) of the conventional applications (Newcombe et al., 2011b; Izadi et al., 2011; PCL, 2011), each voxel contains a pair of *distance value* (D_i) and *accumulated weight* (W_i), describing the truncated distance value to the closest surface and the weight for this value, respectively. This data structure enables averaging of the captured depth data, influencing the voxel model after i frames. For the *(i+1)*th depth frame, the model is updated by the corresponding pair of distance value (d_{i+1}) and weight (w_{i+1}) for voxel x, using the following two equations:

$$D_{i+1}(x) = \frac{W_i(x)D_i(x) + w_{i+1}(x)d_{i+1}(x)}{W_i(x) + w_{i+1}(x)} \ , \quad (1)$$

$$W_{i+1}(x) = W_i(x) + w_{i+1}(x) \ . \quad (2)$$

Parameter d_i is the calculated distance value for voxel x based on the corresponding newly sensed valid depth point, and w_i is the weight of the depth d_i. Depth d_i is integrated into the corresponding voxel x based on Equation (1). The weight for voxel x is accumulated in W_i according to Equation (2). Choosing $w_{i+1} = 1$ for each valid point found in the $(i+1)$th frame, results in simple averaging over time. Unfortunately, according to our experiments, the constant value for the weight affects the synthetic model updating process in the following way. The objects in the model located close to the sensor ($\leq$ 1.5 m) are created properly, while the objects located at a $\geq$ 2.5 m distance are being melted, significantly deformed or even completely destroyed.

2.2 Distance-related weight Definition and Updating Strategies

In this paper, we utilize the weight definition which has been already proposed in (Javan Hemmat et al., 2014a; Javan Hemmat et al., 2014b) to guarantee the assignment of higher weights to points in the scene located on closer distances to the sensor. Based on the sensor features and scene characteristics, there is a valid range for depth data, defined between a maximum and a minimum distance, d_{max} and d_{min}, respectively. In addition, the weight is bounded between 0 and a maximum weight W_{max}. The following equation defines a weight based on the distance to the sensor, specifying

$$weight_{depth\_point}(x) = \left(\frac{\frac{1}{d(x)^2} - \frac{1}{d_{max}^2}}{\frac{1}{d_{min}^2} - \frac{1}{d_{max}^2}} \right) * W_{max} \ . \quad (3)$$

For each depth value x with distance $d(x)$ in the valid range between d_{min} and d_{max}, the corresponding weight is mapped to a value between 0 and W_{max}.

In conventional TSDF model implementations, the model is straightforwardly updated with constant-weight strategy (weight value is unity). The weight definition, introduced by Equation (3), enables us to distinguish between closer and further distances. Therefore, we can exploit this weight definition to intelligently update the TSDF model. The intelligent update prevents more accurate values being overwritten by less accurate data. This updating strategy guarantees that the synthetic 3D model is updated by the most accurate data available during the update process.

Distance-Aware (DA) updating Method

Each voxel value in the synthetic 3D model is updated based on a straightforward intelligent rule: *"if a voxel value has already been updated by a distance value with a higher weight, never update it again by a depth distance with a lower weight"*. The DA updating method is formulated as:

$$Flag(v,x) = weight_{new}(x) \geq r\% \times weight_{LMU}(v), \quad (4)$$

$$Update(v,x) = \begin{cases} \text{Integrate } x \text{ into } v & if\,(Flag(v,x)), \\ \text{Discard } x, \text{ keep } v & \text{otherwise.} \end{cases} \quad (5)$$

To make the updating method more robust to noise, there is a tolerance range, r indicating a percentage with $0 \leq r \leq 100$. Since the distance values are compared to $r\%$ of the last maximum updated weight, the distance values less than the last maximum updated weight are therefore integrated into the synthetic 3D model. This method integrates the distance values close to the last maximum updated weight affected by noise.

Due to intrinsics of the conventional TSDF implementation, the DA method suffers from fast saturation of accumulated weight value, which limits proper truncated-distance averaging over a long sensing process. To eliminate this constraint, the DASS method has been introduced as below (Javan Hemmat et al., 2014a; Javan Hemmat et al., 2014b).

Distance-Aware Slow-Saturation (DASS) updating Method

The DASS method performs similar to the DA method, except for the weight accumulation. The DASS uses the weight definition of Equation (3) for the $Update(v,x)$ function to conditionally update the synthetic 3D model, similar to the DA method. However, in contrast with the DA method, the DASS uses unity for the new weight w_{i+1}, to calculate the weight accumulation value W_{i+1}. This solution of $w_{i+1} = 1$ in the DASS method prevents the fast saturation of the accumulated weight value, while the $Update(v,x)$ function ensures an intelligent updating process.

3 EXPERIMENTS

3.1 Implementation

To implement the proposed updating methods, we have exploited the original framework of the open source KinFu implementation from the Points Cloud Library (PCL, 2011). We have reused the original structure and only inserted the new definitions and updating algorithms as discussed above.

3.2 Dataset

For datasets, we have chosen three outdoor statues located in our campus and recorded them by a stereo camera during daylight, which would not be possible with the Kinect sensor. We have generated depth maps from the obtained stereo data and adopted the maps to the Kinect depth format. Figure 1 shows the statues and their corresponding depth frames.

3.3 Evaluation Approach

Due to absence of ground-truth models, the quality evaluation of models obtained from the original KinFu, DA, and DASS methods is purely based on visual assessment. It is known that in the KinFu algorithm, a reset of the 3D modeling occurs in case of large errors in the pose-estimation algorithm. The pose-estimation accuracy is mutually dependent on the quality of reconstructed model. Besides the qualitative issues, for a quantitative evaluation, we compute the number of resets and also the endurance interval length before the first reset occurs during the 3D reconstruction process, as an indicator of the pose-estimation accuracy.

4 ANALYSIS AND DISCUSSION

According to the results shown in Figure 2, the most interesting finding is that the KinFu-based 3D reconstruction algorithm works well on the outdoor data provided by a stereo camera, which is nearly impossible to achieve with the Kinect sensor data during daylight time.

Figures 2.B-D show that for all datasets, the DA and DASS methods are able to preserve the model and avoid deformation, in contrast with the original KinFu algorithm. The original KinFu easily degrades the model by overwriting the more accurate data with less accurate data. This also proves that the stereo depth data is clearly sensitive to the distance.

Comparing the DA and DASS methods, especially for the third dataset (swordfish), we can conclude the DASS method provides a more smooth and accurate 3D model. This can be explained by a fast saturation of accumulated weights in the DA method, which prevents proper temporal averaging of the depth data.

As a quantitative evaluation illustrated in Table 1, the average number of resets caused by the DASS method is significantly lower than the corresponding numbers from the DA and original KinFu methods. The reason for this reduction is the direct correlation between the model quality and the pose-estimation accuracy (Javan Hemmat et al., 2014b).

Another quantitative metric for model quality is the endurance of an algorithm until the first reset occurrence. For this metric, the DASS algorithm shows a higher performance and is able to sustain more than twice longer than the original KinFu algorithm.

An interesting finding which has revealed during the process, is the level of continuity of the stereo depth data over different surfaces. In contrast with the Kinect as a depth sensor based on the IR-projection, the depth frames extracted from stereo data could be less continuous on featureless surfaces, while the Kinect is able to prepare continuous data as far as the surface is not black or shiny. On the other hand, for

Table 1: Pose-estimation accuracy: comparison of the number of resets and endurance of the DA, DASS and KinFu methods. Higher reset rates indicate lower accuracy in the pose estimation. The pose-estimation accuracy is mutually dependent on the 3D model quality.

Method	**metric**	**Statue 1**	**Statue 2**	**Statue 3**	**Average**
KinFu	number of reset(s)	3	1	1	**1.67**
	endurance (frames)	1,811	2,996	544	**1,783.67**
DA	number of reset(s)	3	1	1	**1.67**
	endurance (frames)	2,195	2,995	544	**1,911.33**
	improvement (%)	21.20	-0.03	0.00	**7.06**
DASS	number of reset(s)	0	0	1	**0.33**
	endurance (frames)	5,604	3,801	544	**3,316.33**
	improvement (%)	209.44	26.87	0.00	**78.77**

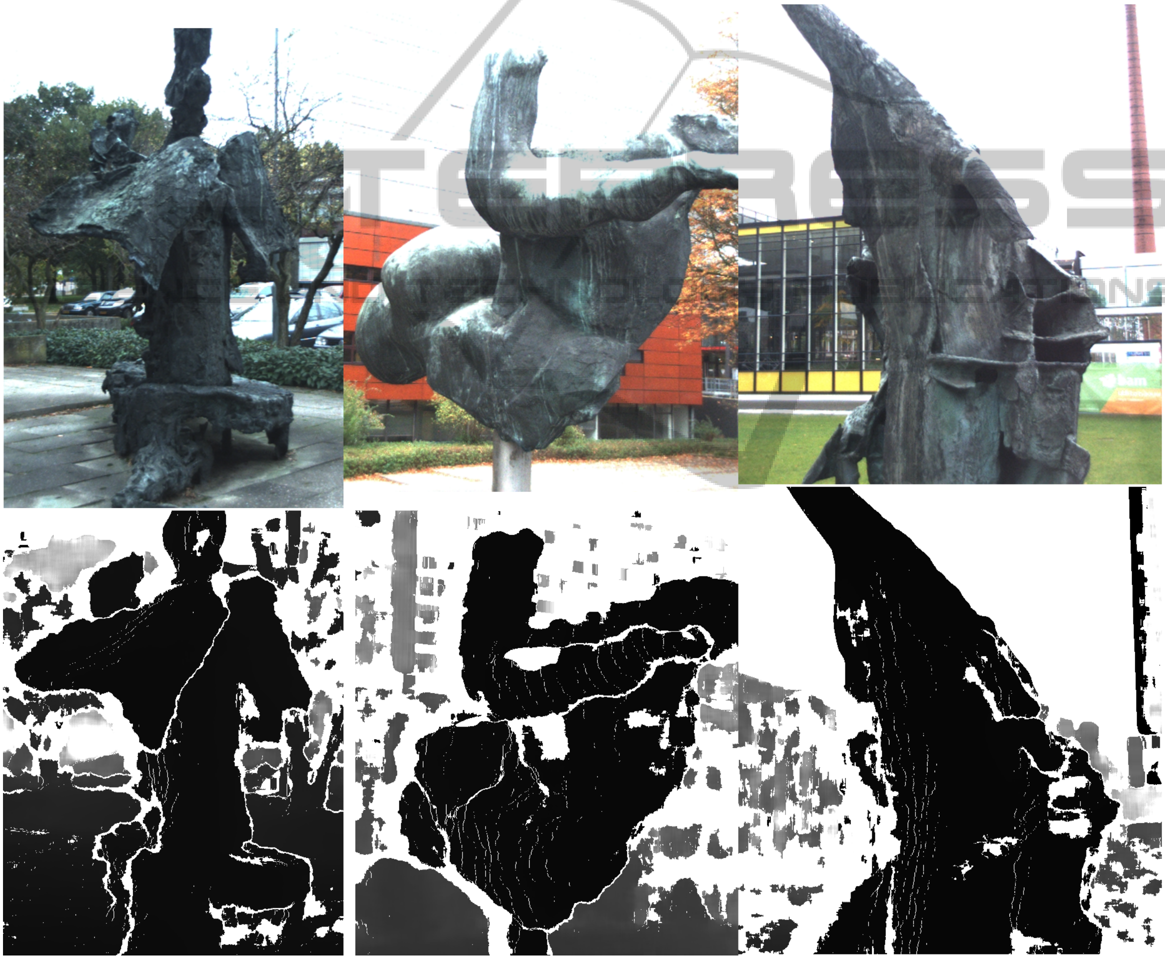

Figure 1: Snapshots of the modern art statues and their corresponding depth frame extracted from stereo data. The depth frames depict the closer depth information as darker points.

the surfaces with sufficient visual features that absorb or distract the IR projected patterns of the Kinect, the stereo camera provides more continuous and smooth depth data. This motivates a hybrid-sensing approach, where both the Kinect and stereo sensors are deployed in a single system setup, to increase reconstruction robustness of heterogeneous scenes.

5 CONCLUSIONS

We have experimented with the deployment of the KinFu algorithm onto the depth input data obtained from stereo sensors to enable 3D reconstruction of outdoor scenes during daylight, which is an impossible task with the Kinect sensor. We have evaluated

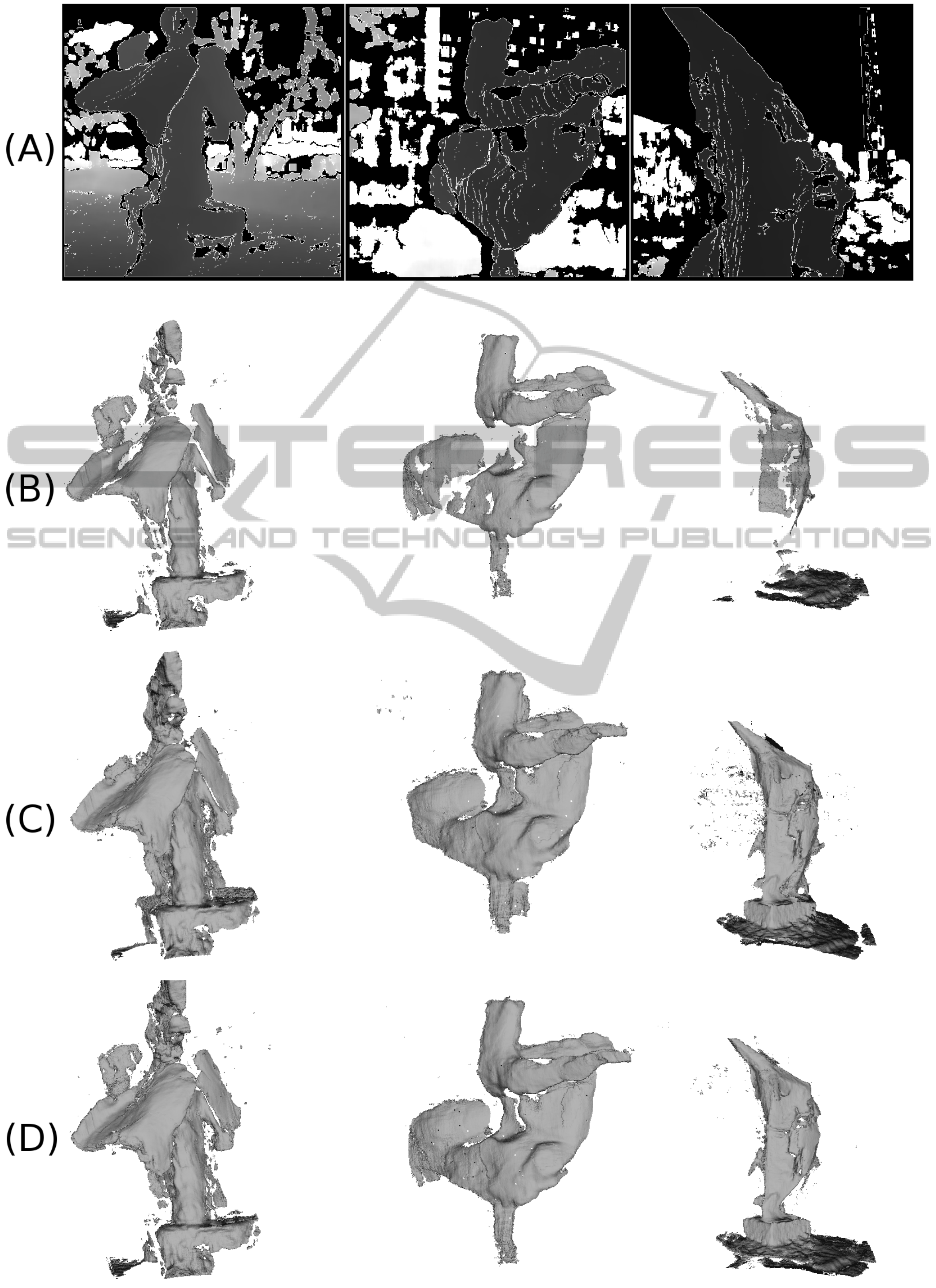

Figure 2: Mesh snapshots obtained by different weighting strategies. Each column shows the result for one of the datasets. (A) shows a converted depth image in the format of the standard Kinect. (B), (C), and (D) illustrate the reconstruction results from the original KinFu, DA, and DASS methods, respectively. Note the model degradation in (B) compared to (C) and (D).

the quality of 3D models reconstructed by the original KinFu algorithm in comparison with our distance-aware DA and DASS methods. The experiments have revealed that the input from stereo sensors is valid and sufficient for KinFu-based algorithms, resulting in an appropriate reconstruction of outdoor scenes. We have also shown that by replacing the original KinFu weighting strategy by distance-aware weighting strategies, we obtain 3D models from stereo data with higher quality and more accurate pose-estimation values. In our experiments, the new strategies increase the endurance of the reconstruction process with a factor of two or more.

Comparing the depth data obtained from the Kinect and stereo sensors, we have found that the stereo camera is able to provide more continuous depth data for scenes with sufficient visual features that interfere the IR patterns of the Kinect sensor, such as black or shiny surfaces. Alternatively, the Kinect can provide more continuous depth data for the surfaces with insufficient amount of visual features or featureless surfaces, where stereo cameras are unable to extract any depth information.

For future work, we plan experiments on finding the optimal hybrid system capable of working in different environments in terms of the quantity and quality of visual and 3D features and intelligently fusing the resulting data from depth sensor and stereo camera, based on the scene configuration and features.

ACKNOWLEDGEMENTS

This research has been performed within the PANORAMA project, co-funded by grants from Belgium, Italy, France, the Netherlands, the United Kingdom, and the ENIAC Joint Undertaking.

REFERENCES

Asus. Xtion-PRO. http://www.asus.com/Multimedia/Xtion_PRO/.

Bondarev, E., Heredia, F., Favier, R., Ma, L., and de With, P. (2013). On photo-realistic 3d reconstruction of large-scale and arbitrary-shaped environments. In *Consumer Communications and Networking Conference (CCNC), 2013 IEEE*, pages 621–624.

Chow, J., Ang, K., Lichti, D., and Teskey, W. (2012). Performance analysis of a low-cost triangulation-based 3D camera: Microsoft kinect system. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B5, 2012 XXII ISPRS Congress*, Melbourne, Australia.

Curless, B. and Levoy, M. (1996). A volumetric method for building complex models from range images. In *CM SIGGRAPH Conf. Proceedings, pp. 303–312 (1996).*

Engelhard, N., Endres, F., Hess, J., Strum, J., and Burgard, W. (2011). Real-time 3d visual slam with a hand-held camera. In *RGB-D Workshop on 3D Perception in Robotics, European Robotics Forum.*

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., and Fitzgibbon, A. (2011). Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 559–568, New York, NY, USA. ACM.

Javan Hemmat, H., Bondarev, E., and de With, P. (2014a). Exploring distance-aware weighting strategies for accurate reconstruction of voxel-based 3d synthetic models. In *MMM'14 Proceedings of Multi-Media Modelling 2014 (will be presented) (2014).*

Javan Hemmat, H., Bondarev, E., Dubbelman, G., and de With, P. (2014b). Improved icp-based pose estimation by distance-aware 3d mapping. In *VISAPP'14, Proceedings of 9th International Conference on Computer Vision Theory and Application (will be presented) 2014.*

Khoshelham, K. (2011). Accuracy analysis of kinect depth data. In *ISPRS workshop laser scanning / ed. by D.D. Lichti and A.F. Habib., IAPRS XXXVIII-5/W12 (2011).*

Khoshelham, K. and Elberink, S. O. (2012). Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454.

Kinect, M. http://www.xbox.com/en-us/kinect/.

Newcombe, R., Lovegrove, S., and Davison, A. (2011a). Dense tracking and mapping in real-time. In *ICCV'11, IEEE International Conference on Computer Vision, Spain.*

Newcombe, R. A., Davison, A. J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D., and Fitzgibbon, A. (2011b). Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 127–136.

PCL (2011). Kinectfusion implementation in the PCL. http://svn.pointclouds.org/pcl/trunk/.

PrimeSense. http://www.primesense.com/.

Steinbrucker, F., Sturm, J., and Cremers, D. (2011). Real-time visual odometry from dense rgb-d images. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 719–722.

Whelan, T., Johannsson, H., Kaess, M., Leonard, J., and McDonald, J. (2012a). Robust tracking for real-time dense rgb-d mapping with kintinuous. In *MIT technical report, MIT-CSAIL-TR-2012-031.*

Whelan, T., Johannsson, H., Kaess, M., Leonard, J., and McDonald, J. (2013). Robust real-time visual odometry for dense RGB-D mapping. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Karlsruhe, Germany.

Whelan, T., Kaess, M., Fallon, M., Johannsson, H., Leonard, J., and McDonald, J. (2012b). Kintinuous: Spatially extended kinectfusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras, Sydney, Australia.*