

The role of RGB-D benchmark datasets: an overview

Kai Berger

Oxford e-Research Centre

7 Keble Road, Oxford

kai.berger@oerc.ox.ac.uk

Abstract

The advent of the Microsoft Kinect three years ago stimulated not only the computer vision community for new algorithms and setups to address well-known problems in the community but also sparked the launch of several new benchmark datasets to which future algorithms can be compared to. This review of the literature and industry developments concludes that the current RGB-D benchmark datasets can be useful to determine the accuracy of a variety of applications of a single or multiple RGB-D sensors.

1. Introduction

The commercial success of the Microsoft Kinect [27] in November 2010 sparked a multitude of significant research papers in the computer vision community. The Microsoft Kinect was originally designed as a motion sensing input device for the gaming console Microsoft XBOX 360 by tracking the player's motions. As structured light sensor, the Kinect emits a defined light spot pattern, which was first patented by PrimeSense. Users can access the Kinect data streams via USB 2.0 with the help of OpenKinect's *libfreenect*. The main advantage of Kinect capturing setups over conventional time-of-flight (ToF) setups is that the cost is only a fraction of the usual ToF-setups, which makes experimenting with one or many Kinects very convenient.

The projected spot pattern, used for computing the depth maps, is generated as follows: an infrared laser, projects a defined pattern at $850nm$ onto all surfaces of the scene facing the sensor in the frustum. The diffuse reflection of the pattern in the scene is captured by a camera, which has its infrared filter removed. An onboard circuit computes the disparity for each 9×9 subpattern by computing the distance to their default positions for an image of a default scene (this is likely to be a wall parallel to the sensor at a defined distance of 3m). The disparity values are mapped to distance values in meters. This technique has been introduced by PrimeSense.

The project pattern is a texture of 211×165 spot positions. 3861 spot positions are brighter, the rest is assumed dark. That pattern is replicated in a 3×3 pattern to broaden the field of view. The central spot of the pattern appears brighter than all the other spots and no two bright spots are adjacent. The pattern looks quasi-random but in fact is the same for all cameras. Thus, one device can compute the depth map from the emitted pattern of another adjacent device, if its own laser is obstructed. It is assumed that the depth computation follows a block search approach, i.e. the integrated circuit looks the corresponding position in the neighbourhood of the original subpattern position and computes depth values from the distance to the original position. The visual tact rate can be computed from a central distinguishable subpattern, where a horizontal line alternates bright and dark spots. As several brighter spots are visible over the complete pattern, it can be assumed, that distortions might be calculated from their location information in the received image.

Located within the Kinect device, there is one RGB camera, that operates at 30Hz with a resolution of 640×480 pixels or 15Hz with a resolution of 1280×1024 pixels, and one IR camera, that operates at 30Hz with a resolution of 640×480 pixels or 10Hz with a resolution of 1280×1024 pixels. In the IR view it can be noticed, that visible light is captured with the chip at a small intensity range as well, and that a Bayer pattern shows when zooming in. It is assumed, that a diffractive element splits up the laser light before it traverses a mechanical grid with occluding material applied to the spots that are designed to be dark in the projection. For completeness reasons it should be noted, that the Kinect also bears a microphone array and an accelerometer. The Kinect can be tilted on command via USB. Most capturing environments presented in this report are indoors, such that daylight with intensity in the IR spectrum does not oversaturate the recorded image.

Together with Asus Xtion and PrimeSense depth acquisition has become significantly easier. Thanks to accurate depth data, currently published papers could present a broad range of RGB-D setups addressing well-known problems

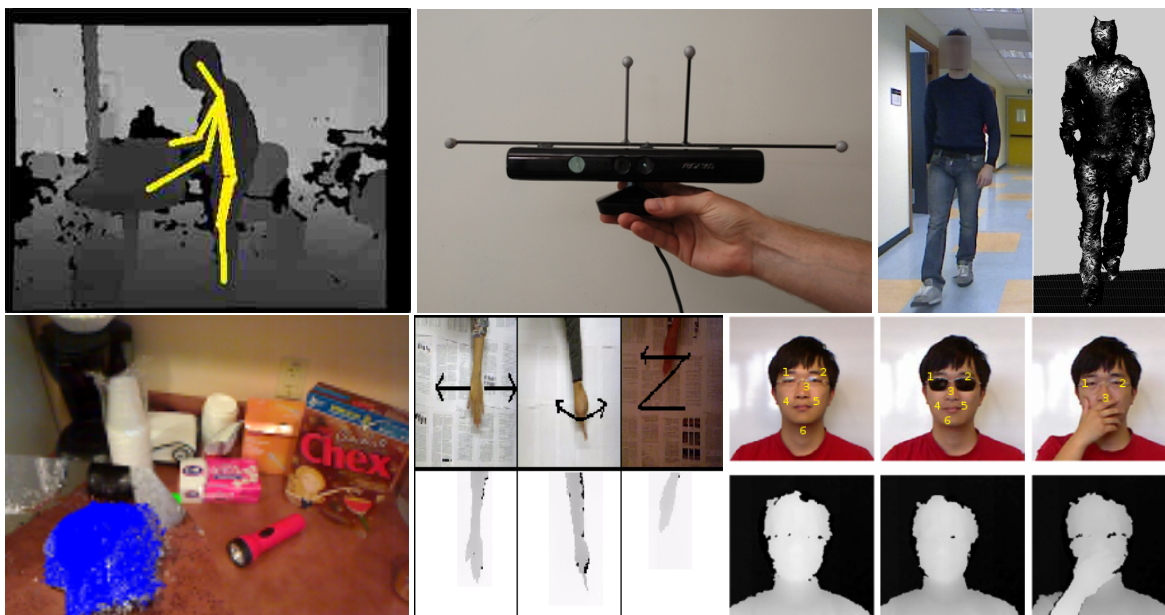


Figure 1. A collage of the variety of benchmark datasets that are currently publicly available. Top left: depth images with annotated motion (reproduced from [39]), top middle: external tracking of Kinect pose with markers (reproduced from [37]), top right: tight mesh and skeleton alongside rgb-data (reproduced from Barbosa *et al.* [5]). Bottom left: depth images with objects annotated (reproduced from [16]), bottom middle: depth data with annotated hand movements (reproduced from [21]), bottom right: face capturings in RGB-D stream anoted (reproduced from Huynh *et al.* [13]).

in computer vision in which the Microsoft Kinect ranging from SLAM [10, 12, 19, 17, 35, 11] over 3d reconstruction [2, 33, 38, 32, 1] over realtime face [18] and hand [30] tracking to motion capturing and gait analysis [34, 41, 8, 7, 4]. The course of the research over the past years also required some datasets captured with the Microsoft Kinect or a similar RGB-D sensor to be made publicly available for comparison.

In the following sections we provide an overview over the significant benchmarks that are currently publicly available for comparison, Fig. 1. A tabular overview about the corresponding publications for each dataset can found in Table 1.

1.1. Method Of Comparison

This overview paper does not provide new research findings but it attempts to provide an overview over the diverse set of benchmarks that are publicly available for comparison of RGB-D based algorithms. The findings are summarized in an overview table, Table 1 and compared for main distinguishable criteria. The table is sorted alphabetically for each research field, i.e. *SLAM*, Sect. 3 and *Object Recognition*, Sect. 4. We evaluated if the accelerometer of the Kinect was used (third column), if the data were annotated and which type of ground truth has been made available (fourth column). Finally we provided the link to the datasets (fifth columns). We tested the accessibility in the middle of August. Some datasets may require login data, which

however can be acquired by contacting the corresponding authors (instructions were published on the corresponding website in that case). In Sect. 5 we provide a critical view onto the diversity of the publicly available datasets and phrase suggestions for extending the state of the art in benchmarks. Statistics about the volume and impact of each dataset is provided in Fig. 2.

2. Annotation for Ground Truth retrieval

Most datasets exceed a feasible size to be handled by a single user for annotation. Hence, with the increasing popularity of internet freelance websites, most publications presented in this report have relied on Mechanical Turk, e.g. [15], for robust annotation of the datasets. Some rely on additional sensors to provide the ground truth, e.g. for the camera pose at a given frame [37, 36]. A sophisticated approach transforms the labeling in another space: instead of letting the user annotate in image space, the static scene captured with a moving Kinect is reconstructed in 3d and annotated in a 3d graphics tool once, e.g. [16]. The annotated point clouds are then simply reprojected into the input stream using the camera pose for the Kinect sensor at each frame.

3. SLAM

Highly accurate depth data are necessary for 3D reconstruction and simultaneous reconstruction and simulta-

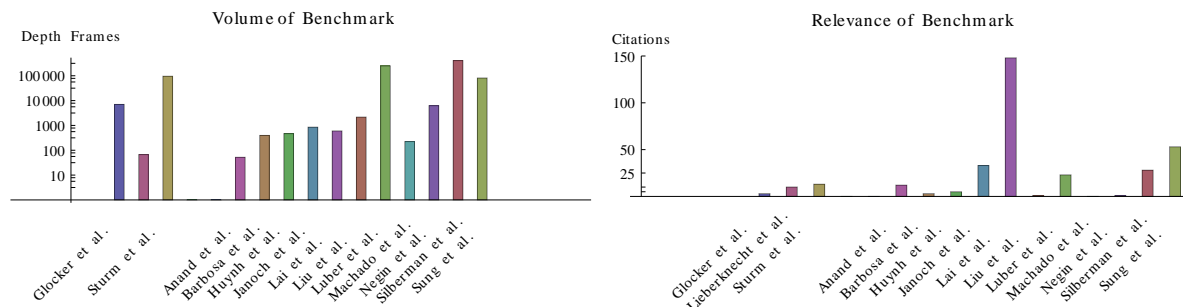


Figure 2. Left: This semi-logarithmic bar chart depicts the size of each published dataset in terms of absolute depth-images. The dataset presented by Silberman *et al.* [28] bears the most input images. Right: This chart depicts the impact of each published dataset in the community. It is sorted alphabetically for each research field. The work by Lai *et al.* [16] has been considered most in the community.

neous localization and mapping (SLAM) applications, although the requirements for mapping or localization can differ within the applicational context. It can be seen, that accuracy and the running time/framerates trade each other off. The Kinect is the first device that provides fast data acquisition at acceptable accuracy. In their work Sturm *et al.* [37, 36] release a 50 GB dataset consisting of 39 RGB-D sequences captured with the Microsoft Kinect including the recorded accelerometer data with the intention to test SLAM algorithms on the input data. The authors provide ground truth via external per frame pose estimation of the Kinect within a global reference framework, which has been computed from the capturing of markers that have been attached to the Kinect beforehand. They used a MotionAnalysis capturing system at 100 Hz. Lieberknecht *et al.* [20] create also a benchmark for localisation and provide video data, from which the RGB and depth data can be extracted. However, they do not provide a dataset that contains annotations or additional data, e.g. accelerometric data. Glocker *et al.* [9] provide a dataset captured with a moving camera and use KinectFusion to generate the 3d scene and the camera path as ground truth for the benchmark. They provide seven different scenes including RGB, depth and pose data in a txt-file.

4. Object Recognition

Based on the Kinect's realtime output of accurate depth maps, it became possible to reconstruct 3D objects with the Kinect, e.g. by moving the sensor around the acquired object. For example, Tam and his colleagues [40] register point clouds captured with the Kinect to each other. Lai *et al.* [16] present an annotated dataset containing visual and depth images of 300 physically distinct objects ranging from fruits to tools. Their dataset was captured with the Primesense prototype and a Firewire RGB-camera from Pointgrey. Their approach to labeling the objects in the input sequences is somewhat innovative: they reconstruct the 3d scene from the moving RGB-D sensor setup while keeping track of its position over time. The objects of in-

terest are then labeled once in the 3d scene by hand and then backprojected into the input streams. Liu *et al.* [21] present a dataset for gesture recognition where 2160 hand gesture sequences of 6 persons are captured with the Microsoft Kinect. The annotated dataset differentiates 10 hand gestures: circle (clockwise), triangle (anti-clockwise), up-down, right-left, wave, Z, cross, comehere, turnaround. As the Microsoft Kinect remains fixed during acquisition there is no additional accelerometric data in the dataset. Negin *et al.* [29] provide a dataset of human body movements represented by 3D positions of skeletal joints. As the Kinect sensor remained fixed, no accelerometric data is available, but the authors provide the complete trackign results gained from applying the Microsoft Kinect SDK to the RGB-D data as the ground truth for their benchmark. In the dataset 15 people conduct 10 different exercises. Barbosa *et al.* [5] capture 79 persons first for a distinctive signature, e.g. in a defined pose, and then in regular motion, e.g. walking across a floor. They provide both skeleton fits and .ply meshes alongside the RGB-D data. The goal of their dataset is to reidentify different humans captured with the Kinect. The humans may change their movement patterns or their clothes in between recordings. Machado *et al.* [24] record several objects and models with the Kinect camera and let them annotate by human observers. The meshes are presented in various formats with the task to identify the object from the recorded shape. Luber *et al.* [22] present a pedestrian dataset captured with three Kinects which are placed such that their viewing cones do not interfere. The dataset is annotated in that the position of each pedestrian is bounded by a rectangle in the input views. Their dataset contains of walking and standing pedestrians seen from different orientations and with different levels of occlusions. Silberman *et al.* [28] present a dataset consisting of 1449 labeled pairs of aligned RGB and depth images captured in indoor environments, such as bathrooms, basements, bedrooms, kitchens and playrooms. It includes the accelerometric data for each frame and also features a toolbox implemented in matlab that includes useful functions for manipulating the data and

labels. Anand *et al.* [3] captured several indoor environments and labeled the depth data. They also present in bag files the output of RGBDSLAM for each scene, e.g. for each timestamp a transform-matrix for that frame that transforms the camera from the first frame accordingly. Janoch *et al.* [15] show a large dataset annotated with the help of Amazon's Mechanical Turk consisting of indoor environment items like chairs, monitors, cups, bottles, bowls, keyboards, mouses or phones. They do not provide additional accelerometer data. Dataset consisting of faces of 52 people (14 females, 38 males) captured with the Microsoft Kinect has been presented by Huynh *et al.* [13]. The faces are captured in nine different conditions (neutral face, smile, mouth open, face in left profile, face in right profile, partial occlusion of face parts, changing lighting conditions). They do not include the accelerometric data. Defined landmark points were manually identified in the input images. In their work about motion recognition Sung *et al.* [39] provide depthmaps and skeletons for four subjects (two male, two female, one left-handed) who were asked to perform different high-level activities, like making cereal, arranging objects or having a meal. The activities are label and sub-classified for movements like reaching, opening, placing, or scrubbing.

5. Shortcomings

The authors believe that, although there is already quite a remarkable amount of publicly available datasets based on capturings conducted with the Kinect, certain aspects in use of the sensor seem to be underrepresented. While already one paper is published [25] that aims to extend the depth reconstruction capabilities from IR input stream data, a coherent dataset containing the IR data and additional ground truth depth information, e.g. from scene calibration or stereo, is missing. Also, arbitrary mesh reconstruction is in the datasets currently considered as byproduct of SLAM algorithms, Sect. 3, such that estimates with the accuracy of a few millimeters to a centimeter seem sufficient. However, recently publications have emerged to employ one or many Kinects for the accurate reconstruction of objects, e.g. based on depth, a combination of depth and texture cues in the RGB stream [26] or from IR input stream [31]. The reconstructed objects in these setups need explicitly not necessarily be purely opaque [23, 6]. A ground truth dataset with a high-resolution laser scan alongside input frames from Kinect (depth, RGB and IR) with a pose reconstruction of the sensor position would be highly desirable.

6. Conclusion

In this state-of the art report we have provided an overview over the publicly available datasets generated for benchmark with the Microsoft Kinect. Several approaches,

ranging from a steady single Kinect capturing setup over a moving Kinect in the scene to capturing setups that include multiple Kinects, have been discussed. The applicational context varied between SLAM, motion capturing and recognition. We have also phrased a critical view onto the diversity of current datasets with suggestions for extending the state of the art in benchmarks. With the deployment of the new *Kinect One* in the near future the authors assume that in the next years the amount of publicly available benchmark datasets will increase significantly.

References

- [1] N. Ahmed. A system for 360 acquisition and 3d animation reconstruction using multiple rgb-d cameras.
- [2] D. S. Alexiadis, G. Kordelas, K. C. Apostolakis, J. D. Agapito, J. Vegas, E. Izquierdo, and P. Daras. Reconstruction for 3d immersive virtual environments. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on*, pages 1–4. IEEE, 2012.
- [3] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, 32(1):19–34, 2013.
- [4] S. Asteriadis, A. Chatzitofis, D. Zarpalas, D. S. Alexiadis, and P. Daras. Estimating human motion from multiple kinect sensors. In *Proceedings of the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, page 3. ACM, 2013.
- [5] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *Computer Vision—ECCV 2012. Workshops and Demonstrations*, pages 433–442. Springer, 2012.
- [6] K. Berger, K. Ruhl, M. Albers, Y. Schroder, A. Scholz, J. Kokemuller, S. Guthe, and M. Magnor. The capturing of turbulent gas flows using multiple kinects. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1108–1113. IEEE, 2011.
- [7] K. Berger, K. Ruhl, C. Brümmer, Y. Schröder, A. Scholz, and M. Magnor. Markerless motion capture using multiple color-depth sensors. In *Proc. Vision, Modeling and Visualization (VMV)*, volume 2011, page 3, 2011.
- [8] A. L. Fuhrmann, J. Kretz, and P. Burwik. Multi sensor tracking for live sound transformation.
- [9] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi. Real-time rgb-d camera relocalization. In *International Symposium on Mixed and Augmented Reality*. Springer, 2013.
- [10] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *the 12th International Symposium on Experimental Robotics (ISER)*, volume 20, pages 22–25, 2010.
- [11] D. Herrera C, J. Kannala, and J. Heikkilä. Accurate and practical calibration of a depth and color camera pair. In *Computer Analysis of Images and Patterns*, pages 437–445. Springer, 2011.

- [12] G. Hu, S. Huang, L. Zhao, A. Alempijevic, and G. Disanayake. A robust rgb-d slam algorithm. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, 2012.
- [13] T. Huynh, R. Min, and J.-L. Dugelay. An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In *Computer Vision-ACCV 2012 Workshops*, pages 133–145. Springer, 2013.
- [14] S. Izadi, R. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In *ACM SIGGRAPH 2011 Talks*, page 23. ACM, 2011.
- [15] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013.
- [16] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.
- [17] T. Lee, S. Lim, S. Lee, S. An, and S. Oh. Indoor mapping using planes extracted from noisy rgb-d sensors. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, 2012.
- [18] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo. Kinect identity: Technology and experience. *Computer*, 44(4):94–96, 2011.
- [19] S. Lieberknecht, A. Huber, S. Ilic, and S. Benhimane. Rgb-d camera-based parallel tracking and meshing. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 147–155. IEEE, 2011.
- [20] S. Lieberknecht, A. Huber, S. Ilic, and S. Benhimane. RGB-D camera-based parallel tracking and meshing. In *ISMAR*, 2011.
- [21] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2013.
- [22] M. Luber, L. Spinello, and K. O. Arras. People tracking in rgb-d data with on-line boosted target models. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 3844–3849. IEEE, 2011.
- [23] I. Lysenkov, V. Eruhimov, and G. R. Bradski. Recognition and pose estimation of rigid transparent objects with a kinect sensor. In *Robotics: Science and Systems*, 2012.
- [24] J. Machado and A. Ferreira. Retrieval of objects captured with low-cost depth-sensing cameras. In *SHREC2013*. Springer, 2013.
- [25] M. Martinez and R. Stiefelhagen. Kinect unleashed: Getting control over high resolution depth maps.
- [26] D. Miao, J. Fu, Y. Lu, S. Li, and C. W. Chen. Texture-assisted kinect depth inpainting. In *Circuits and Systems (IS-CAS), 2012 IEEE International Symposium on*, pages 604–607. IEEE, 2012.
- [27] Microsoft News Center. Microsoft press release.
- [28] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012.
- [29] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, and A. Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In *Image Analysis and Recognition*, pages 648–657. Springer, 2013.
- [30] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. *BMVC*, Aug, 2, 2011.
- [31] T.-H. Ou-Yang, M.-L. Tsai, C.-T. Yen, and T.-T. Lin. An infrared range camera-based approach for three-dimensional locomotion tracking and pose reconstruction in a rodent. *Journal of neuroscience methods*, 201(1):116–123, 2011.
- [32] A. Pancham, N. Tlale, and G. Bright. Mapping and tracking of moving objects in dynamic environments. 2012.
- [33] N. Rafibakhsh, J. Gong, M. K. Siddiqui, C. Gordon, and H. F. Lee. Analysis of xbox kinect sensor data for use on construction sites: Depth accuracy and sensor interference assessment. In *Constitution Research Congress*, pages 848–857, 2012.
- [34] A. Santhanam, D. Low, and P. Kupelian. Th-c-brc-11: 3d tracking of interfraction and intrafraction head and neck anatomy during radiotherapy using multiple kinect sensors. *Medical Physics*, 38:3858, 2011.
- [35] J. Smisek, M. Jancosek, and T. Pajdla. 3d with kinect. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1154–1160. IEEE, 2011.
- [36] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the IEEE Int. Conf. on Intelligent Robot Systems (IROS)*, pages 573–580, 2012.
- [37] J. Sturm, S. Magnenat, N. Engelhard, F. Pomerleau, F. Colas, W. Burgard, D. Cremers, and R. Siegwart. Towards a benchmark for rgb-d slam evaluation. In *Proc. of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conf.(RSS), Los Angeles, USA*, volume 2, page 3, 2011.
- [38] L. Sumar and A. Bainbridge-Smith. Feasibility of fast image processing using multiple kinect cameras on a portable platform. *Department of Electrical and Computer Engineering, Univ. Canterbury, New Zealand*.
- [39] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgb-d images. In *Plan, Activity, and Intent Recognition*, 2011.
- [40] G. Tam, Z.-Q. Cheng, Y.-K. Lai, F. Langbein, Y. Liu, A. Marshall, R. Martin, X.-F. Sun, and P. Rosin. Registration of 3d point clouds and meshes: A survey from rigid to non-rigid. 2012.
- [41] A. D. Wilson and H. Benko. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 273–282. ACM, 2010.

Author	Intended Application	Datasize	Accelerometer Data	Annotated	Link
Glocker <i>et al.</i> [9]	SLAM	151MB	No	Camera Path generated with KinectFusion [14]	http://research.microsoft.com/en-us/projects/7-scenes/
Lieberknecht <i>et al.</i> [20]	SLAM	≈ 100 kB	No	No	https://www.dropbox.com/sh/1ky-hns6s1xpbmzw/RQKaYqdp7B/videos
Sturm <i>et al.</i> [37]	SLAM	50GB	Yes	Ground truth pose via external markers tracked with motion capturing system	https://cvpr.in.tum.de/research/datasets/rgbd-dataset
Anand <i>et al.</i> [3]	Object Recognition	≈ 7.6 GB	Yes	Annotated Depth images	http://pr.cs.cornell.edu/sceneunderstanding/data/data.php
Barbosa <i>et al.</i> [5]	Object Recognition	456 MB	No	Skeleton and Meshes	http://www.iit.it/en/datasets/rgbdid.html
Huynh <i>et al.</i> [13]	Object Recognition	no information	No	Faces labeled in input data	http://rgb-d.eurecom.fr/
Janoch <i>et al.</i> [15]	Object Recognition	793 MB	No	Objects labeled in input data	http://www.eecs.berkeley.edu/~allie/VOCB3DO.zip
Lai <i>et al.</i> [16]	Object Recognition	84GB	No	Objects labeled in input data	http://www.cs.washington.edu/rgbd-dataset
Liu <i>et al.</i> [21]	Object Recognition	≈ 1 GB	No	Hand gestures labeled in input data	http://lshao.staff.shef.ac.uk/data/Sheffield-Kinect-Gesture.htm
Luber <i>et al.</i> [22]	Object Recognition	2 GB	No	Pedestrians labeled in input data	http://www.informatik.uni-freiburg.de/~spinello/sw/rgbd_people_unihall.tar.gz
Machado <i>et al.</i> [24]	Object Recognition	24.5MB	No	Objects Labeled	https://dl.dropbox.com/u/4151663/OR/Dataset/test%20set.zip
Negin <i>et al.</i> [29]	Object Recognition	142GB	No	Motion Files containing the tracked joints	http://vpa.sabanciuniv.edu/databases/WorkoutSU-10/MinimalDataset.rar
Silberman <i>et al.</i> [28]	Object Recognition	428GB	Yes	Labeled Depth Dataset	http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
Sung <i>et al.</i> [39]	Object Recognition	≈ 13.8 GB	No	Skeleton and activity/reachability labels	http://pr.cs.cornell.edu/humanactivities/data.php

Table 1. Overview table for the benchmark datasets that are publicly available. We compared properties like data size (third row), the availability of the accelerometric data (fourth row) and the amount of annotation for ground truth (fifth row). For all datasets we listed the link under which they are publicly available. However, some datasets may require the request for login data.