

## Chapter 2

# A State of the Art Report on Multiple RGB-D Sensor Research and on Publicly Available RGB-D Datasets

Kai Berger

**Abstract** That the Microsoft Kinect, an RGB-D sensor, transformed the gaming and end consumer sector has been anticipated by the developers. That it also impacted in rigorous computer vision research has probably been a surprise to the whole community. Shortly before the commercial deployment of its successor, Kinect One, the research literature fills with resumes and state-of-the art papers to summarize the development over the past 3 years. This chapter describes significant research projects which have built on sensing setups that include two or more RGB-D sensors in one scene and on RGB-D datasets captured with them which were made publicly available.

## 2.1 Introduction

With the release of the Microsoft Kinect in November 2010, Microsoft predicted a significant change in the use of gaming devices in the end consumer market. After a preview at the E3 game convention in the Windows Media Centre Environment, the selling in North America started at November 4, 2010 and up to today more than 24 million units have been sold. With the release of an open-source SDK named *libfreenect* by Hèctor Martìn that enables streaming both the depth and the RGB or the raw infrared images via USB the attention of young researchers to use the Microsoft Kinect sensor for their imaging and reconstruction applications has gained. It was possible to stream 1, 200 × 960 RGB and IR images at a frame rate of 30 Hz alongside computed depth estimates of the scene at a lower resolution. The IR image featured the projected infrared pattern generated with an 830 nm laser diode, which is distinctive and the same for each device. Shortly thereafter the proceedings

---

K. Berger (✉)  
INRIA Rennes, Bretagne Atlantique, Campus Universitaire de Beaulieu, 35042 Rennes Cedex,  
France  
e-mail: kai.berger@inria.fr

and journals in the community included papers describing a broad range of setups addressing well-known problems in computer vision in which the Microsoft RGB-D sensor was employed. The projects ranged from simultaneous localization and mapping (SLAM) over 3D reconstruction over realtime face and hand tracking to motion capturing and gait analysis. Counter-intuitively researchers became soon interested in addressing the question if it is possible to employ several Microsoft Kinects, i.e. RGB-D sensors, in one setup—and if so, how to mitigate interference errors in order to enhance the signal. This idea is mainly counter-intuitive due to the fact, the each device projects the same pattern at the same wavelength into the scene. Thus, one would expect that the confusion in processing the raw IR-data rises quickly with the amount of sensors installed in a scene, Fig. 2.1. In the following sections I give an overview over several research projects published in the proceedings and journals of the computer vision community that successfully overcome this preconception and highlight their challenges as well as the benefit of each multiple RGB-D sensor setup. In the second half I list the most prominent datasets, that are publicly available, which were generated with RGB-D sensor setups. A tabular overview about addressed papers is found in Table 2.1. This overview over the state-of-the-art differs from other Kinect-related overview reports in that it does neither include an in-depth evaluation of Time-Of-Flight sensors [20] nor a detailed introduction into the functionality of the sensor algorithm itself [14] nor does it focus on work capturing faces and gestures only [56]. Instead it provides an overview over multiple Kinect setups (Sects. 2.2–2.6) and publicly available databases generated with one or multiple Kinects (Sects. 2.7–2.11).

## 2.2 Multiple Kinect-Setups: Method of Comparison

As this chapter is a state-of-the art report it explicitly provides no new research contribution. Instead it shall be read as an overview and introduction to the work that has been conducted in the subfield of multiple Kinect research. I want to provide



**Fig. 2.1** A simple scene (*left*) captured with the depth camera of one (*middle*) and multiple concurrently projecting kinects (*right*). The interference of more than one Kinect pattern results in degradations in the captured depth image (*white* pixels denote invalid depth values). This state of the art report lists significant papers that implemented setups albeit interference issues or to specifically address and overcome these issues. Reproduced from Schroeder et al. [44]

**Table 2.1** An overview over different publications including multiple RGB-D sensors

Authors	Context	Number of RGB-D sensors in setup	Accuracy	Calibrated
Asteriadis et al. [4]	Motion estimation	3	Not specified	Yes
Berger et al. [8]	Motion estimation	4	Reprojection error of 1.7 px	Yes
Fuhrmann et al. [12]	Motion estimation	3	Deviation of 2–3 cm	Yes
Hossny et al. [15]	Motion estimation	2	Not specified	Yes (the authors provide a new autocalibration algorithm)
Santhanam et al. [40]	Motion estimation	4	Deviation of 3 mm	Yes
Wilson [53]	Motion estimation	3	Not specified	Yes
Ye et al. [54]	Motion estimation	3	Not specified	No
Zhang et al. [55]	Motion estimation	2	Deviation of 20 cm	Yes
Alexiadis et al. [2]	Mesh reconstruction	4	Reprojection error of 0.8 px	Yes
Berger et al. [7] and Berger et al. [6]	Mesh reconstruction	3	Not specified	Yes
Macknoja et al. [28]	Mesh reconstruction	5	Deviation of 2.5 cm at 3 m distance	Yes
Lo et al. [24]	Mesh reconstruction	2	Not specified	Yes
Nakamura [32]	Mesh reconstruction	2	Deviation of 3 % at 90° spacing	No
Nakazawa et al. [33]	Mesh reconstruction	4	Not specified	Yes
Ahmed [1]	Mesh reconstruction	6	Not specified	Yes
Olesen et al. [36]	Mesh reconstruction	3	60 % inlier at 8 px textlet spacing	Yes

(continued)

Table 2.1 (continued)

Authors	Context	Number of RGB-D sensors in setup	Accuracy	Calibrated
Pancham et al. [38]	Mesh reconstruction	2+	Not specified	No
Rafibakhsh et al. [39]	Mesh reconstruction	2	Deviation of 3.49 cm	Yes
Sumar et al. [47]	Mesh reconstruction	2	Reprojection error of 5 px	No
Tong et al. [51]	Mesh reconstruction	3	Biometrical measure deviation 1.6–6.2 cm	Yes
Wang et al. [52]	Mesh reconstruction	2	Not specified	Yes
Caon et al. [10]	Recognition	3	Not specified	Yes
Satta et al. [42]	Recognition	2	Not specified	No
Satyavolu et al. [43]	Recognition	5	Deviation of 3 cm	Yes
Saputra et al. [41]	Recognition	2	Deviation of 10 cm	Yes
Susanto et al. [49]	Recognition	5	Deviation of 13 cm	Yes
Butler et al. [9]	Interference	2 and 3	Deviation of up to 3 cm	Yes
Faion et al. [11]	Interference	4	Deviation of 21 mm	Yes
Kainz et al. [19]	Interference	8	Not specified	Yes
Maimone and Fuchs [29]	Interference	6	Deviation of 2 mm	No
Schroeder et al. [44] and Berger et al. [8]	Interference	4	Reprojection error of 1.7 px	Yes

The table lists for each publication the amount of employed sensors, the context of application, the accuracy and whether the sensors were calibrated to a common world space. Note, that the specification of accuracy varies with the context of application between the mean deviation of a reconstructed 3D position from the original position in meters and the reprojection error in pixels or percentage into the camera

a comparative table, Table 2.2 for the reader to have a quick overview of examined papers and their properties. The table is sorted alphabetically for each research field, i.e. *Multiple RGB-D sensor Setups for Motion Estimation*, Sect. 2.3, *Multiple RGB-D sensor Setups for Reconstruction*, Sect. 2.4, *Multiple RGB-D sensor Setups for Recognition and Tracking*, Sect. 2.5, and *Interference in Multiple RGB-D sensor Setups*, Sect. 2.6. I compared the amount of Kinects installed in each capturing environment (third column), and stated where the sources were available the measured accuracy of the capturings. As the statements were not unified, I have to provide them in different units to adhere to the source text. A slightly more detailed description is given at the table caption. Finally I state if the capturing setup was externally calibrated to a common worldspace, usually performed with a checkerboard or moving a marker around the scene.

### 2.3 Multiple RGB-D Sensor Setups for Motion Estimation

Santhanam et al. [40] describe a system to track neck and head movements with four calibrated Kinects. Three Kinects are tracking the patient's anatomy contour in depth and RGB streams while the fourth camera detects the face of the patient. The detected face region is used to guide the contour detection in the other three views. The detected contours are then finally merged to to a 3D estimate of the pose of the anatomy. The authors claim a precision of 3 mm at the expected 30 Hz. Wilson and Benko [53] use three PrimeSense depth cameras which stream at  $320 \times 240$  px resolution and 30 Hz for human interaction with an augmented reality table. They compare input depth image streams against background depth images for each depth camera captured when the room is empty to segment out the human user. While the authors do not specify the accuracy, e.g. between the projected area and the captured area comprised by a hand, they claim to robustly track all user actions in 10 cm volume above the table. The depth cameras were placed next to each other and slanted such that each camera captures a different angle in the room. However their viewing cones may have overlapped. Fuhrmann et al. [12] have employed a stage setup with three Kinects for musical performances. They calibrated the cameras, which were observing the same  $3 \times 3 \times 3$  m<sup>3</sup> interaction volume from different angles, for each stage performance. The tracking via *OpenNI* suffered only from latency between interframe capturing times. The sensors were employed such that they did not interfere destructively. Berger et al. [8] employ four Kinect sensors in a small  $3 \times 3 \times 3$  m<sup>3</sup> room to mitigate shortcomings in the motion capturing capabilities of a single Kinect, Fig. 2.2 (left). To overcome depth map degradation through interfering patterns they introduced external hardware shutters. The idea was further evaluated by Zhang et al. [55] who basically performed the same capturing only with two Kinect cameras. Interference issues were circumvented by placing them opposite each other and assuming that the human actor acts as a separation surface between both projection cones. The authors claim a tracking accuracy of 20 cm. Their processing algorithm limits the original capturing framerate of 30–15 Hz. Asteriadis et al. [4] included a treadmill to

**Table 2.2** Overview table for the benchmark datasets that are publicly available

Authors	Intended application	Datasize	Accelerometer data	Annotated	Link
Glocker et al. [13]	SLAM	151 MB	No	Camera path generated with kinectfusion [17]	<a href="http://research.microsoft.com/en-us/projects/7-scenes/">http://research.microsoft.com/en-us/projects/7-scenes/</a>
Lieberknecht et al. [22]	SLAM	≈100 kB	No	No	<a href="http://www.dropbox.com/sh/1kyhns6s1xpbmzw/RQKaYqdp7B/videos">http://www.dropbox.com/sh/1kyhns6s1xpbmzw/RQKaYqdp7B/videos</a>
Sturm et al. [46]	SLAM	50 GB	Yes	Ground truth pose via external markers tracked with motion capturing system	<a href="http://cvpr.in.tum.de/research/datasets/rgbd-dataset">http://cvpr.in.tum.de/research/datasets/rgbd-dataset</a>
Anand et al. [3]	Object recognition	≈7.6 GB	Yes	Annotated depth images	<a href="http://pr.cs.cornell.edu/sceneunderstanding/data/data.php">http://pr.cs.cornell.edu/sceneunderstanding/data/data.php</a>
Barbosa et al. [5]	Object recognition	456 MB	No	Skeleton and meshes	<a href="http://www.iiit.it/en/datasets/rgbdid.html">http://www.iiit.it/en/datasets/rgbdid.html</a>
Huynh et al. [16]	Object recognition	No information	No	Faces labeled in input data	<a href="http://rgb-d.eurecom.fr/">http://rgb-d.eurecom.fr/</a>
Janoch et al. [18]	Object recognition	793 MB	No	Objects labeled in input data	<a href="http://www.eecs.berkeley.edu/~allie/VOCB3DO.zip">http://www.eecs.berkeley.edu/~allie/VOCB3DO.zip</a>
Lai et al. [21]	Object recognition	84 GB	No	Objects labeled in input data	<a href="http://www.cs.washington.edu/rgbd-dataset">http://www.cs.washington.edu/rgbd-dataset</a>
Liu and Shao [23]	Object recognition	≈1 GB	No	Hand gestures labeled in input data	<a href="http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm">http://lshao.staff.shef.ac.uk/data/SheffieldKinectGesture.htm</a>

(continued)

Table 2.2 (continued)

Authors	Intended application	Datasize	Accelerometer data	Annotated	Link
Luber et al. [25]	Object recognition	2 GB	No	Pedestrians labeled in input data	<a href="http://www.informatik.uni-freiburg.de/~spinello/sw/rgbd_people_unihall.tar.gz">http://www.informatik.uni-freiburg.de/~spinello/sw/rgbd_people_unihall.tar.gz</a>
Machado and Ferreira [27]	Object recognition	24.5 MB	No	Objects labeled	<a href="http://dl.dropbox.com/u/4151663/OR/Dataset/test%20set.zip">http://dl.dropbox.com/u/4151663/OR/Dataset/test%20set.zip</a>
Negin et al. [35]	Object recognition	142 GB	No	Motion files containing the tracked joints	<a href="http://vpa.sabanciuniv.edu/databases/WorkoutSU-10/MinimalDataset.rar">http://vpa.sabanciuniv.edu/databases/WorkoutSU-10/MinimalDataset.rar</a>
Silberman et al. [34]	Object recognition	428 GB	Yes	Labeled dataset	<a href="http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html">http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html</a>
Sung et al. [48]	Object recognition	≈13.8 GB	No	Skeleton and activity/reachability labels	<a href="http://pr.cs.cornell.edu/humanactivities/data.php">http://pr.cs.cornell.edu/humanactivities/data.php</a>

I compared properties like data size (third row), the availability of the accelerometer data (fourth row) and the amount of annotation for ground truth (fifth row). For all datasets I listed the link under which they are publicly available. However, some datasets may require the request for login data

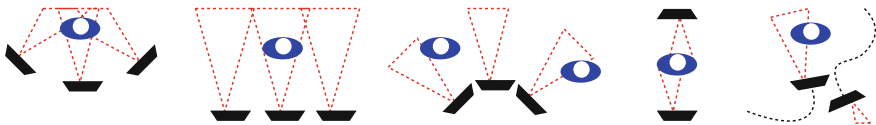
simulate partially occluded motion for three calibrated Kinect sensors placed evenly in a quarter arc around the treadmill. Using a Fuzzy Inference system they were able to robustly map the human motion. Although they do neither state reprojection errors nor deviations from a reconstructed mesh they provide figures that the human motion could be fitted by a skeleton in up to 95 % of the recorded frames. An approach to analyse facial motion with two Kinects is presented by Hossny et al. [15]. They also provide a smart algorithm to automatically calibrate one Kinect to another based on one rotation to zero angular positions. The processing of the depth maps to the face is done with geometric features that outperform conventional Haar features. They propose to overcome interference difficulties with mutually rotated polarization filters but do not state figures about the reprojection error. Very recently, Ye et al. [54] provided a solution for capturing human motion with multiple moving Kinects. In their setup, three Kinects were employed.

## 2.4 Multiple RGB-D Sensor Setups for Reconstruction

Alexiadis et al. [2] use four Kinect devices to reconstruct a single, full 3D textured mesh of a human body from their depth data in realtime. The authors claim that the re-projection error is less than 0.8 pixels. In a merging step redundant triangles are clipped. Object boundary noise is removed with a distance-to-background map. Rafibakhsh et al. [39] analyse construction site scenarios with two Kinects and exhaustively search for optimal placement and angles, concluding that the two sensors should not directly face each other. In their calibrated sensor setup they found a scene accuracy of 3.49 cm. Sumar et al. [47] test the sensor interference for two uncalibrated Kinect sensors in an indoor environment. They found, that in a marker tracking task, where the markers are less than 3 m from the Kinect the error follows a Gaussian distribution and does not deviate more than 5 pixels from the true centre of the marker. In ongoing work Pancham et al. [38] mount Kinects atop mobile robots which move in an overcast outdoor environment in order to segment out moving objects from static scenery. In that context the Kinect is used for differentiation between moving and stationary objects, and for map construction of the environment. They however do not state the accuracy of the reconstructed scene in relation to the amount of Kinects employed. In a very interesting approach to enable HDR scene capturing Lo et al. [24] place two Kinects atop each other and equip one with a polarized neutral density filter. This results in accurate depth values for regions that would have been overexposed in an unaltered Kinect capturing (The exposure difference between both IR images is roughly 1 EV apart). They recognise the fact that interference might occur but did not quantitatively evaluate that for their setup. However, the reconstructed scenes bear more complete meshes under headlight than with a single LDR capturing. Berger et al. [7] show in their paper the feasibility to use three Kinects concurrently in a convergent setup for capturing non-opaque surfaces like the interface between flowing propane gas in air. It is noteworthy that, although the projectors are masked such that they project on mutually disjoint surface areas,



the projection patterns do not interfere destructively with each other while passing through the gas volume. Their approach has been altered such that an evaluation based only on the high resolution IR stream is possible as well [6]. Olesen et al. [36] show a system that involves up to three calibrated Kinects for textlet reconstruction. They evaluate different angular settings for the multiple sensors but interestingly conclude that the orientation does not significantly improve the capturing quality. In industrial applications Macknoja et al. [28] place three Kinects on a straight line next to each other while a fourth and a fifth Kinect are placed to the left and right respectively in a convergent manner to provide a calibrated capturing volume with a side length of 7 m in total, Fig. 2.2 (middle). Small projecting volumes overlap while objects like cars are captured. The authors state a depth error of about 2.5 cm at 3 m distance. Wang et al. [52] present work where two calibrated Kinects' depth maps are fused to reconstruct arbitrary scene content. The cameras are spaced 30 cm apart and the viewing axes converge towards the scene centre. Inaccuracies due to interference are handled in software by applying a his work Ahmed [1] provides a scene reconstruction mainly of human bodies captured from 6 calibrated Kinects. He deliberately excludes interference analysis from the discussion but mentioned temporal drift if software synchronization is omitted. Interference issues are also neglected by Nakazawa et al. [33] who placed four calibrated Kinects at the four corners of a capturing room, but rotated them by  $90^\circ$  such that they would capture a greater vertical range and a smaller horizontal range each. They concentrate on aligning depth data captured asynchronously by applying a temporal calibration by providing depth data at certain time instants. Tong et al. [51] reconstruct the human body from a setup consisting of three Kinects mounted on two poles at different heights. The subject is placed on a spinning turntable in the center of the poles. The deviation in different biometrical measures is stated to be in 1.6–6.2 cm. In their work Nakamura [32] place two Kinects in different angles between  $10^\circ$  and  $180^\circ$  from each other around the scene. The Kinects are not calibrated to a common world space but placed at a fixed distance to the scene centre. In an evaluation of the mean reprojection error for the varying angles they find that a spacing of  $180^\circ$  between each Kinect results in the smallest error while a spacing of  $120^\circ$  results in the largest error, Fig. 2.2 (right). The Kinects do not project into each others sensor due to the scene content.



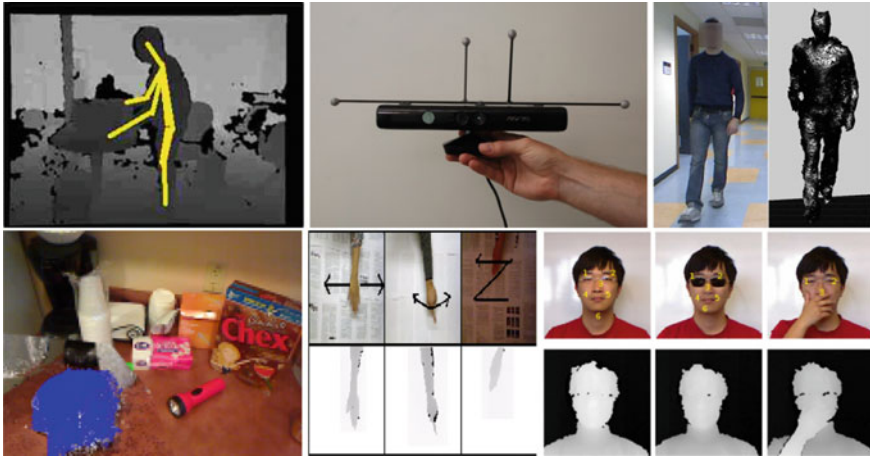
**Fig. 2.2** Five typical capturing setups featuring multiple Kinects. Multiple Kinects are evenly placed in a virtual circle around the scene centre (*first*), e.g. [7, 8, 19, 33, 36, 43, 52], multiple Kinects are in line to capture a volume with a large side length (*second*), e.g. [10, 24, 28, 29, 41], multiple Kinects juxtaposed and facing away from each other (*third*), e.g. [53], and two Kinects face each other, but are occluded by the scene content (*fourth*), e.g. [32]. Very recently work has been conducted with multiple uncalibrated moving Kinects (*fifth*), e.g. [38, 54]

## 2.5 Multiple RGB-D Sensor Setups for Recognition and Tracking

Satta et al. [42] present research to recognize and track people in an indoor environment surveyed by two Kinects relying on a combination of RGB texture and depth information. It has to be noted, though, that the Kinects were installed facing away from each other. Hence, they did not directly project into each other's viewing frustra. Interference is not discussed further. Satyavolu et al. [43] describe an experimental setup that consists of 5 Kinects. One camera was used for tracking IR markers attached to a box, 4 others (evenly distributed around the scene centre) simulated interference/noise. The authors report that the Kinect deviated by 3 cm on an average from the actual position. Caon et al. [10] present an approach for tracking gestures based on three calibrated Kinects placed in a 45° angle. They varied different configurations between the three Kameras and although they did not state figures about the depth or tracking accuracy they do list the amount of invalid depth pixels for each configuration. Susanto et al. [49] present an approach to detect objects from their shape and depth profile generated when captured from several calibrated Kinects and state that there is no degrading interference noticeable due to the fact the the Kinects are placed at wide angles from each other. Although the paper focus on the success rate of the recognition they briefly state that the setup might show depth discrepancies of up to 13 cm. The tracking of humans in a room has been shown by Saputra et al. [41] who placed two calibrated Kinects at 5 m distance next to each other. Although the projection cones do not interfere with each other, the authors provide a detection error of human position of 10 cm.

## 2.6 Interference in Multiple RGB-D Sensor Setups

Following the work of Berger et al. [8], where external hardware shutters are used for mitigating interference between concurrently projecting sensors as described in detail by Schroeder et al. [44], Maimone and Fuchs [29] introduce motion platforms that pitch each Kinect with the Kinect that the own structured light pattern remains crisp in the IR stream while the other patterns appear blurred due to the angular motion of the camera. The depth map is realigned with the recorded egomotion from the inertial sensors included in the Kinect. It is noteworthy that they also managed to deblur the RGB-image using the Lucy-Richardson method. In a more generic approach Butler et al. [9] vibrate the camera arbitrarily. In a rather invasive approach Faion et al. [11] manage to toggle the projector subsystem to perform measurements similar to Schroeder et al. [44]. They use Bayesian state estimator to intelligently schedule which sensor is to be selected for the next time frame. Their maximal reconstruction error denotes 21 mm. Kainz et al. [19] describe an elaborate setup for eight Kinects mounted on vibrating rods and one freely moving Kinect suitable for various applications, such as motion capturing and reconstruction. All vibrating rods

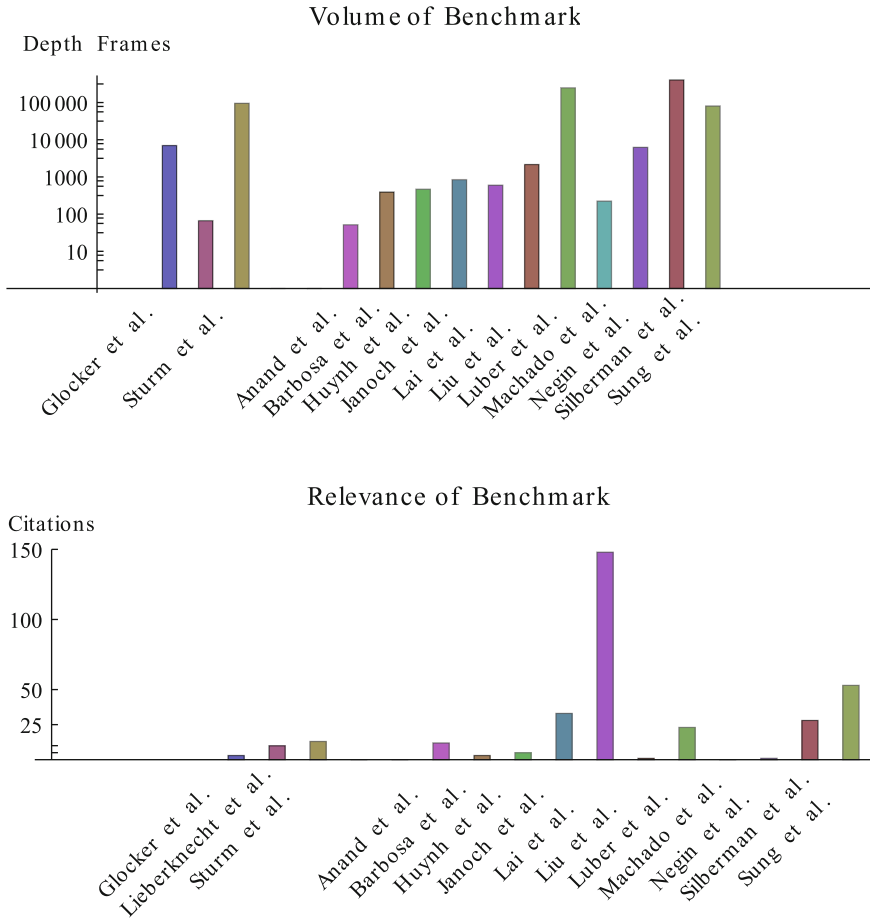


**Fig. 2.3** A collage of the variety of benchmark datasets that are currently publicly available. *Top left* depth images with annotated motion (reproduced from [48]), *top middle* external tracking of kinect pose with markers (reproduced from [46]), *top right* tight mesh and skeleton alongside RGB-data (reproduced from Barbosa et al. [5]). *Bottom left* depth images with objects annotated (reproduced from [21]), *bottom middle* depth data with annotated hand movements (reproduced from [23]), *bottom right* face captursings in RGB-D stream anotated (reproduced from Huynh et al. [16])

were administered by a parallel circuit at slightly different frequencies. They do not give a quantitative analysis of the reconstruction error but provide qualitative figures of the reconstructed mesh (Fig. 2.3).

## 2.7 RGB-D Datasets: Method of Comparison

In this part of the chapter it is attempted 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 2.2 and compared for main distinguishable criteria. The table is sorted alphabetically for each research field, i.e. *SLAM*, Sect. 2.9 and *Object Recognition*, Sect. 2.10. I 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 I provided the link to the datasets (fifth columns). I 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. 2.11 I 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. 2.4.



**Fig. 2.4** *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. [34] 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. [21] has been considered most in the community

## 2.8 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. [18], 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 [45, 46]. 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. [21]. The annotated point clouds are then simply reprojected into the input stream using the camera pose for the Kinect sensor at each frame.

## 2.9 SLAM

Highly accurate depth data are necessary for 3D reconstruction and simultaneous reconstruction and 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. [45, 46] 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. [22] 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. [13] 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.

## 2.10 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 [50] register point clouds captured with the Kinect to each other. Lai et al. [21] 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 interest are then labeled once in the 3D scene by hand and then backprojected into the input streams. Liu and Shao [23] present a dataset for gesture recognition where 2,160 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. [35] 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 tracking 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 and Ferreira [27] 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. [25] 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. [34] present a dataset consisting of 1,449 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. [18] 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, mice 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. [16]. 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. [48] 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 subclassified for movements like reaching, opening, placing, or scrubbing.

## 2.11 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 [30] 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. 2.9, 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 [31] or from IR input stream [37]. The reconstructed objects in these setups need explicitly not necessary be purely opaque [7, 26]. 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.

## 2.12 Conclusion

In this chapter I have shown that, counter-intuitively, it is possible to use several Kinects in one capturing setup. Although each device projects the same pattern at the same wavelength into the scene and consequently contributes to confusion in processing the raw IR-data, several approaches, ranging from hardware fixes over intelligent software algorithms for mitigation to placing the Kinects such that the scene content acts as an occluding surface between each projection cone, have been discussed. The applicational context varied between motion capturing, the original purpose of the Kinect sensor, over scene reconstruction to tracking and recognition. Furthermore, I 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. I 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. It has to be evaluated, though, if setups with multiple sensors in one capturing scenario are possible, but the authors predict that in the next years there will still be challenges for multiple RGB-D sensors relying on the emission of light to be addressed by the community

## References

1. Ahmed N (2012) A system for 360 acquisition and 3D animation reconstruction using multiple RGB-D cameras
2. Alexiadis DS, Kordelas G, Apostolakis KC, Agapito JD, Vegas J, Izquierdo E, Daras P (2012) Reconstruction for 3D immersive virtual environments. In: 13th international workshop on image analysis for multimedia interactive services (WIAMIS). IEEE, pp 1–4



3. Anand A, Koppula HS, Joachims T, Saxena A (2013) Contextually guided semantic labeling and search for three-dimensional point clouds. *Int J Robot Res* 32(1):19–34
4. Asteriadis S, Chatzitofis A, Zarpalas D, Alexiadis DS, Daras P (2013) Estimating human motion from multiple kinect sensors. In: *Proceedings of the 6th international conference on computer vision/computer graphics collaboration techniques and applications*. ACM, p 3
5. Barbosa IB, Cristani M, Del Bue A, Bazzani L, Murino V (2012) Re-identification with RGB-D sensors. In: *Computer vision-ECCV 2012. Workshops and demonstrations*. Springer, pp 433–442
6. Berger K, Kastner M, Schroeder Y, Guthe S (2013) Using sparse optical flow for two-phase gas flow capturing with multiple kinects. *Robotics: science and systems 2013 workshop on RGB-D: advanced reasoning with depth cameras*, pp 1–8
7. Berger K, Ruhl K, Albers M, Schroder Y, Scholz A, Kokemuller J, Guthe S, Magnor M (2011) The capturing of turbulent gas flows using multiple kinects. In: *IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, pp 1108–1113
8. Berger K, Ruhl K, Brümmer C, Schröder Y, Scholz A, Magnor M (2011) Markerless motion capture using multiple color-depth sensors. In *Proceedings of vision, modeling and visualization (VMV)*, vol 2011, p 3
9. Butler DA, Izadi S, Hilliges O, Molyneaux D, Hodges S, Kim D (2012) Shake'n'sense: reducing interference for overlapping structured light depth cameras. In: *Proceedings of the 2012 ACM annual conference on human factors in computing systems*. ACM, pp 1933–1936
10. Caon M, Yue Y, Tscherrig J, Mugellini E, Abou Khaled O (2011) Context-aware 3D gesture interaction based on multiple kinects. In: *AMBIENT 2011, the first international conference on ambient computing, applications, services and technologies*, pp 7–12
11. Faion F, Friedberger S, Zea A, Hanebeck UD (2012) Intelligent sensor-scheduling for multi-kinect-tracking. In: *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp 3993–3999
12. Fuhrmann AL, Kretz J, Burwik P (2013) Multi sensor tracking for live sound transformation
13. Glocker B, Izadi S, Shotton J, Criminisi A (2013) Real-time RGB-D camera relocalization. In: *International symposium on mixed and augmented reality*. Springer
14. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review
15. Hossny M, Filippidis D, Abdelrahman W, Zhou H, Fielding M, Mullins J, Wei L, Creighton D, Puri V, Nahavandi S (2012) Low cost multimodal facial recognition via kinect sensors. In: *Proceedings of the land warfare conference (LWC): potent land force for a joint maritime strategy*. Commonwealth of Australia, pp 77–86
16. Huynh T, Min R, Dugelay J-L (2013) An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In: *Computer vision-ACCV 2012 workshops*. Springer, pp 133–145
17. Izadi S, Newcombe R, Kim D, Hilliges O, Molyneaux D, Hodges S, Kohli P, Shotton J, Davison A, Fitzgibbon A (2011) Kinectfusion: real-time dynamic 3D surface reconstruction and interaction. In: *ACM SIGGRAPH 2011 talks*. ACM, p 23
18. Janoch A, Karayev S, Jia Y, Barron JT, Fritz M, Saenko K, Darrell T (2013) A category-level 3D object dataset: putting the kinect to work. In: *Consumer depth cameras for computer vision*. Springer, pp 141–165
19. Kainz B, Hauswiesner S, Reitmayr G, Steinberger M, Grasset R, Gruber L, Veas E, Kalkofen D, Seichter H, Schmalstieg D (2012) Omnikinect: real-time dense volumetric data acquisition and applications. In: *Proceedings of the 18th ACM symposium on virtual reality software and technology*. ACM, pp 25–32
20. Khoshelham K (2011) Accuracy analysis of kinect depth data. In: *ISPRS workshop laser scanning*, vol 38, p 1
21. Lai K, Bo L, Ren X, Fox D (2011) A large-scale hierarchical multi-view RGBD-D object dataset. In: *IEEE international conference on robotics and automation (ICRA)*. IEEE, pp 1817–1824



22. Lieberknecht S, Huber A, Ilic S, Benhimane S (2011) RGB-D camera-based parallel tracking and meshing. In: ISMAR
23. Liu L, Shao L (2013) Learning discriminative representations from RGB-D video data. In: Proceedings of the international joint conference on artificial intelligence (IJCAI)
24. Lo R, Rampersad V, Huang J, Mann S (2013) Three dimensional high dynamic range veillance for 3D range-sensing cameras
25. Luber M, Spinello L, Arras KO (2011) People tracking in RGBD-D data with on-line boosted target models. In: IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 3844–3849
26. Lysenkov I, Eruhimov V, Bradski GR (2012) Recognition and pose estimation of rigid transparent objects with a kinect sensor. In: Robotics: science and systems
27. Machado J, Ferreira A (2013) Retrieval of objects captured with low-cost depth-sensing cameras. In: SHREC2013. Springer
28. Macknoja R, Chávez-Aragón A, Payeur P, Laganière R (2013) Calibration of a network of kinect sensors for robotic inspection over a large workspace. In: Proceedings of the IEEE workshop on robot vision (WoRV 2013)
29. Maimone A, Fuchs H (2012) Reducing interference between multiple structured light depth sensors using motion. In: Virtual reality workshops (VR). IEEE, pp 51–54
30. Martinez M, Stiefelhagen R (2013) Kinect unleashed: getting control over high resolution depth maps
31. Miao D, Fu J, Lu Y, Li S, Chen CW (2012) Texture-assisted kinect depth inpainting. In: IEEE international symposium on circuits and systems (ISCAS). IEEE, pp 604–607
32. Nakamura DALR Multiple 3D data acquisition system setup based on structured lighth technique for immersive videoconferencing applications
33. Nakazawa M, Mitsugami I, Makihara Y, Nakajima H, Habe H, Yamazoe H, Yagi Y (2012) Dynamic scene reconstruction using asynchronous multiple kinects. In: 21st international conference on pattern recognition (ICPR). IEEE, pp 469–472
34. Nathan Silberman PK, Hoiem D, Fergus R (2012) Indoor segmentation and support inference from RGBD images. In: ECCV
35. Negin F, Özdemir F, Akgül CB, Yüksel KA, Erçil A (2013) A decision forest based feature selection framework for action recognition from RGB-depth cameras. In: Image analysis and recognition. Springer, pp 648–657
36. Olesen SM, Lyder S, Kraft D, Krüger N, Jessen JB (2012) Real-time extraction of surface patches with associated uncertainties by means of kinect cameras. *J Real-Time Image Process* 1–14
37. Ou-Yang T-H, Tsai M-L, Yen C-T, Lin T-T (2011) An infrared range camera-based approach for three-dimensional locomotion tracking and pose reconstruction in a rodent. *J Neurosci Methods* 201(1):116–123
38. Panoram A, Tlale N, Bright G (2012) Mapping and tracking of moving objects in dynamic environments
39. Rafibakhsh N, Gong J, Siddiqui MK, Gordon C, Lee HF (2012) Analysis of xbox kinect sensor data for use on construction sites: depth accuracy and sensor interference assessment. In: Constitution research congress, pp 848–857
40. Santhanam A, Low D, Kupelian P (2011) Th-c-brc-11: 3D tracking of interfraction and intrafraction head and neck anatomy during radiotherapy using multiple kinect sensors. *Med Phys* 38:3858
41. Saputra MRU, Putra GD, Santosa PI et al (2012) Indoor human tracking application using multiple depth-cameras. In: International conference on advanced computer science and information systems (ICACSIS). IEEE, pp 307–312
42. Satta R, Pala F, Fumera G, Roli F (2013) Real-time appearance-based person re-identification over multiple kinect TM cameras
43. Satyavolu S, Bruder G, Willemsen P, Steinicke F (2012) Analysis of IR-based virtual reality tracking using multiple kinects. In: Virtual reality workshops (VR). IEEE, pp 149–150

44. Schröder Y, Scholz A, Berger K, Ruhl K, Guthe S, Magnor M (2011) Multiple kinect studies. *Comput Graph*
45. Sturm J, Engelhard N, Endres F, Burgard W, Cremers D (2012) A benchmark for the evaluation of RGB-D slam systems. In: *Proceedings of the IEEE international conference on intelligent robot systems (IROS)*, pp 573–580
46. Sturm J, Magnenat S, Engelhard N, Pomerleau F, Colas F, Burgard W, Cremers D, Siegwart R (2011) Towards a benchmark for RGB-D slam evaluation. In: *Proceedings of the RGB-D workshop on advanced reasoning with depth cameras at robotics: science and systems conference (RSS)*, vol 2. Los Angeles, USA, p 3
47. Sumar L, Bainbridge-Smith A (2014) Feasability of fast image processing using multiple kinect cameras on a portable platform. Department of electrical and computer engineering, University. Canterbury, New Zealand
48. Sung J, Ponce C, Selman B, Saxena A (2011) Human activity detection from RGBD images. In: *plan, activity, and intent recognition*
49. Susanto W, Rohrbach M, Schiele B (2012) 3D object detection with multiple kinects. In: *Computer vision-ECCV 2012. Workshops and demonstrations*. Springer, pp 93–102
50. Tam G, Cheng Z-Q, Lai Y-K, Langbein F, Liu Y, Marshall A, Martin R, Sun X-F, Rosin P (2012) Registration of 3D point clouds and meshes: a survey from rigid to non-rigid
51. Tong J, Zhou J, Liu L, Pan Z, Yan H (2012) Scanning 3D full human bodies using kinects. *IEEE Trans Visual Comput Graph* 18(4):643–650
52. Wang J, Zhang C, Zhu W, Zhang Z, Xiong Z, Chou PA (2012) 3D scene reconstruction by multiple structured-light based commodity depth cameras. In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp 5429–5432
53. Wilson AD, Benko H (2010) 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*. ACM, pp 273–282
54. Ye G, Liu Y, Deng Y, Hasler N, Ji X, Dai Q, Theobalt C (2013) Free-viewpoint video of human actors using multiple handheld kinects. In: *IEEE transactions on cybernetics*
55. Zhang L, Sturm J, Cremers D, Lee D (2012) Real-time human motion tracking using multiple depth cameras. In: *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp 2389–2395
56. Zhang Z (2012) Microsoft kinect sensor and its effect. *IEEE Multimedia* 19(2):4–10

Computer Vision and Machine Learning with RGB-D  
Sensors

Shao, L.; Han, J.; Kohli, P.; Zhang, Z. (Eds.)

2014, X, 316 p. 163 illus., 148 illus. in color., Hardcover

ISBN: 978-3-319-08650-7