

# A State Of the Art Report on Research in Multiple RGB-D sensor Setups

Kai Berger  
Oxford e-Research Centre  
University Of Oxford  
kai.berger@oerc.ox.ac.uk

## 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 particular report describes significant research projects which have built on sensing setups that include two or more RGB-D sensors in one scene.*

## 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 Martín 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 1200x960 RGB and IR images at a framerate of 30Hz alongside computed depth estimates of the scene at a lower resolution. The IR image featured the projected infrared pattern generated with an 830nm laser diode, which is distinctive and the same for each device. Shortly thereafter the proceedings 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 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 ques-

tion 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. 1. In the following sections we 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. A tabular overview about addressed papers is found in Table 1.

### 1.1. Method Of Comparison

As this paper 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. We want to provide a comparative table, Table 1, to have a short index of examined papers and the properties. The table is sorted alphabetically for each research field, i.e. *Multiple RGB-D sensor Setups for Motion Estimation*, Sect. 2, *Multiple RGB-D sensor Setups for Reconstruction*, Sect. 3, *Multiple RGB-D sensor Setups for Recognition and Tracking*, Sect. 4, and *Interference in Multiple RGB-D sensor Setups*, Sect. 5. We 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, we 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 we state if the capturing setup was externally calibrated to a common worldspace, usually performed with a checkerboard or moving a marker around the scene.

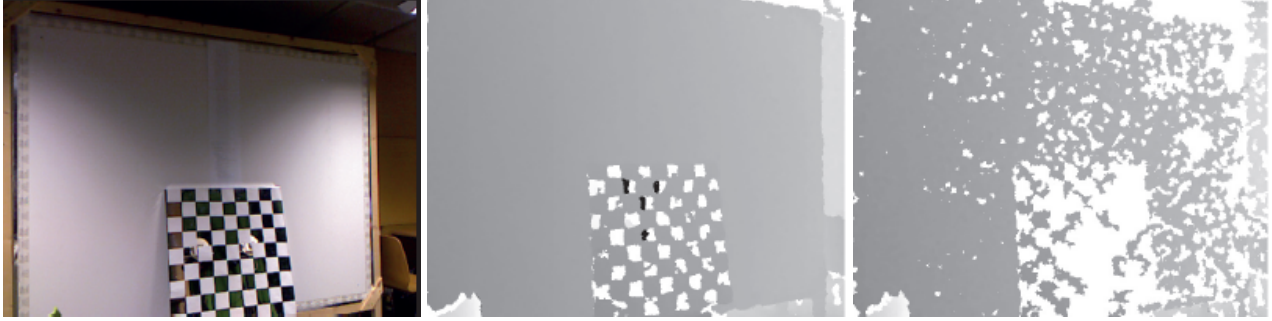


Figure 1. A simple scene (left) captured with the depth camera of one (middle) and multiple concurrently projecting Kinects (right). The interference of more than one Kinect pattern results in degradations in the captured depth image (white pixels denote invalid depth values). This state of the art report lists significant papers that implemented setups albeit interference issues or to specifically address and overcome these issues. Reproduced from Schroeder *et al.* [25].

## 2. Multiple RGB-D sensor Setups for Motion Estimation

Santhanam *et al.* [21] describe a system to track neck and head movements with four calibrated Kinects. Three Kinects are tracking the patients 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  $3mm$  at the expected  $30Hz$ . Wilson *et al.* [29] use three PrimeSense depth cameras which stream at  $320 \times 240px$  resolution and  $30Hz$  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 juxtaposed 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.* [10] have employed a stage setup with three Kinects for musical performances. They calibrated the cameras, which were observing the same  $3 \times 3 \times 3m^3$  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.* [6] employ four Kinect sensors in a small  $3 \times 3 \times 3m^3$  room to mitigate shortcomings in the motion capturing capabilities of a single Kinect, Fig. 2 (left). To overcome depth map degradation through interfering patterns they introduced external hardware shutters. The idea was further evaluated by Zhang *et al.* [31] who basically performed the same capturing only with two Kinect cameras. Interference issues were circum-

vented 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  $20cm$ . Their processing algorithm limits the original capturing framerate of  $30Hz$  to  $15Hz$ . Asteriadis *et al.* [3] included a treadmill to 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.* [11]. They also provide a smart algorithm to automatically calibrate one Kinect to another based one rotation to zero angular positions. The processing of the depth maps to the 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.* [30] provided a solution for capturing human motion with multiple moving Kinects. The details about the number of Kinects used in the setup is currently not known to the authors of this report. Also, no knowledge about figures of Reconstruction errors exists currently.

## 3. 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 reprojection 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.* [20] analyse construction site scenarios with two Kinects and exhaustively search for optimal placement an angles,

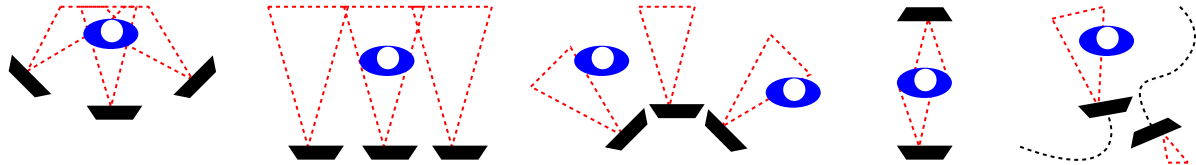


Figure 2. Five typical capturing setups featuring multiple Kinects. Multiple Kinects are evenly placed in a virtual circle around the scene centre (first), e.g. [6, 5, 18, 28, 17, 24, 12], multiple Kinects are in line to capture a volume with a large side length (second), e.g. [14, 13, 8, 15, 22], multiple kinects juxtaposed and facing away from each other (third), e.g. [29], and two Kinects face each other, but are occluded by the scene content (fourth), e.g. [16]. Very recently work has been conducted with multiple uncalibrated moving Kinects (fifth), e.g. [19, 30]

concluding that the two sensors should not directly face each other. In their calibrated sensor setup they found a scene accuracy of  $3.49cm$ . Sumar *et al.* [26] 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 meters 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.* [19] 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.* [13] juxtapose two Kinects atop each other and equip one with a polarized neutral density filter resulting 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.* [5] 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 [4]. Olesen *et al.* [18] 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.* [14] juxtapose 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  $7m$  in total, Fig. 2 (middle). Small projecting volumes overlap while objects like cars are captured. The authors state a depth error of about  $2.5cm$  at  $3m$  distance. Wang *et al.* [28] present work where two calibrated Kinects' depth maps are fused to reconstruct arbitrary scene content. The cameras are spaced  $30cm$  apart and the viewing axes converge towards the scene centre. Inaccuracies due to interference are handled in software by applying a his work Naveed [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.* [17] 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. In their work Nakara *et al.* [16] 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 a spacing of  $120^\circ$  results in the largest error, Fig. 2 (right). The Kinects do not project into each others sensor due to the scene content.

#### 4. Multiple RGB-D sensor Setups for Recognition and Tracking

Satta *et al.* [23] 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.* [24] 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  $3\text{cm}$  on an average from the actual position. Caon *et al.* [8] present an approach for tracking gestures based on three calibrated Kinects placed in a  $45^\circ$  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.* [27] 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  $13\text{cm}$ . The tracking of humans in a room has been shown by Saputra *et al.* [22] who juxtaposed two calibrated Kinects at  $5\text{m}$  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  $10\text{cm}$ .

## 5. Interference in Multiple RGB-D sensor Setups

Following the work of Berger *et al.* [6], where external hardware shutters are used for mitigating interference between concurrently projecting sensors as described in detail by Schroeder *et al.* [25], Maimone and Fuchs [15] 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.* [7] vibrate the camera arbitrarily. In a rather invasive approach Faion *et al.* [9] manage to toggle the projector subsystem to perform measurements similar to Schroeder *et al.* [25]. They use Bayesian state estimator to intelligently schedule which sensor is to be selected for the next time frame. Their maximal reconstruction error denotes  $21\text{mm}$ . Kainz *et al.* [12] 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 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.

## 6. Conclusion

In this state-of the art report we 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. With the advent of Kinect One and the change in underlying technology it will be possible to setup multiple sensors in one capturing scenario more conveniently, 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.

## Appendix

We want to thank the anonymous reviewers for their comments. We thank Yannic Schroeder for providing us with the image material for Figure 1.

## References

- [1] N. Ahmed. A system for 360 acquisition and 3d animation reconstruction using multiple rgb-d cameras. [3](#), [6](#)
- [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. [2](#), [6](#)
- [3] 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. [2](#), [6](#)
- [4] K. Berger, M. Kastner, Y. Schroeder, and S. Guthe. 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*, pages 1–8, June 2013. [3](#), [6](#)
- [5] 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. [3](#), [6](#)
- [6] K. Berger, K. Ruhl, Y. Schroeder, C. Bruemmer, A. Scholz, and M. A. Magnor. Markerless motion capture using multiple color-depth sensors. In *VMV*, pages 317–324, 2011. [2](#), [3](#), [4](#), [6](#)



- [7] D. A. Butler, S. Izadi, O. Hilliges, D. Molyneaux, S. Hodges, and D. Kim. 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*, pages 1933–1936. ACM, 2012. 4, 6
- [8] M. Caon, Y. Yue, J. Tscherrig, E. Mugellini, and O. Abou Khaled. Context-aware 3d gesture interaction based on multiple kinects. In *AMBIENT 2011, The First International Conference on Ambient Computing, Applications, Services and Technologies*, pages 7–12, 2011. 3, 4, 6
- [9] F. Faion, S. Friedberger, A. Zea, and U. D. Hanebeck. Intelligent sensor-scheduling for multi-kinect-tracking. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3993–3999. IEEE, 2012. 4, 6
- [10] A. L. Fuhrmann, J. Kretz, and P. Burwik. Multi sensor tracking for live sound transformation. 2, 6
- [11] M. Hossny, D. Filippidis, W. Abdelrahman, H. Zhou, M. Fielding, J. Mullins, L. Wei, D. Creighton, V. Puri, and S. Nahavandi. Low cost multimodal facial recognition via kinect sensors. In *LWC 2012: Potent land force for a joint maritime strategy: Proceedings of the 2012 Land Warfare Conference*, pages 77–86. Commonwealth of Australia. 2, 6
- [12] B. Kainz, S. Hauswiesner, G. Reitmayr, M. Steinberger, R. Grasset, L. Gruber, E. Veas, D. Kalkofen, H. Seichter, and D. Schmalstieg. Omnikinect: real-time dense volumetric data acquisition and applications. In *Proceedings of the 18th ACM symposium on Virtual reality software and technology*, pages 25–32. ACM, 2012. 3, 4, 6
- [13] R. Lo, V. Rampersad, J. Huang, and S. Mann. Three dimensional high dynamic range veillance for 3d range-sensing cameras. 3, 6
- [14] R. Macknoja, A. Chávez-Aragón, P. Payeur, and R. Laganière. 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)*. 3, 6
- [15] A. Maimone and H. Fuchs. Reducing interference between multiple structured light depth sensors using motion. In *Virtual Reality Workshops (VR), 2012 IEEE*, pages 51–54. IEEE, 2012. 3, 4, 6
- [16] D. d. A. L. R. Nakamura. Multiple 3d data acquisition system setup based on structured ligh technique for immersive videoconferencing applications. 3, 6
- [17] M. Nakazawa, I. Mitsugami, Y. Makiyara, H. Nakajima, H. Habe, H. Yamazoe, and Y. Yagi. Dynamic scene reconstruction using asynchronous multiple kinects. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 469–472. IEEE, 2012. 3, 6
- [18] S. M. Olesen, S. Lyder, D. Kraft, N. Krüger, and J. B. Jessen. Real-time extraction of surface patches with associated uncertainties by means of kinect cameras. *Journal of Real-Time Image Processing*, pages 1–14, 2012. 3, 6
- [19] A. Pancham, N. Tlale, and G. Bright. Mapping and tracking of moving objects in dynamic environments. 2012. 3, 6
- [20] 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. 2, 6
- [21] 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. 2, 6
- [22] M. R. U. Saputra, G. D. Putra, P. I. Santosa, et al. Indoor human tracking application using multiple depth-cameras. In *Advanced Computer Science and Information Systems (ICACSIS), 2012 International Conference on*, pages 307–312. IEEE, 2012. 3, 4, 6
- [23] R. Satta, F. Pala, G. Fumera, and F. Roli. Real-time appearance-based person re-identification over multiple kinect tm cameras. 3, 6
- [24] S. Satyavolu, G. Bruder, P. Willemsen, and F. Steinicke. Analysis of ir-based virtual reality tracking using multiple kinects. In *Virtual Reality Workshops (VR), 2012 IEEE*, pages 149–150. IEEE, 2012. 3, 6
- [25] Y. Schröder, A. Scholz, K. Berger, K. Ruhl, S. Guthe, and M. Magnor. Multiple kinect studies. *Computer Graphics*, 2011. 2, 4, 6
- [26] 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*. 3, 6
- [27] W. Susanto, M. Rohrbach, and B. Schiele. 3d object detection with multiple kinects. In *Computer Vision—ECCV 2012. Workshops and Demonstrations*, pages 93–102. Springer, 2012. 4, 6
- [28] J. Wang, C. Zhang, W. Zhu, Z. Zhang, Z. Xiong, and P. A. Chou. 3d scene reconstruction by multiple structured-light based commodity depth cameras. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5429–5432. IEEE, 2012. 3, 6
- [29] 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. 2, 3, 6
- [30] G. Ye, Y. Liu, Y. Deng, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Free-viewpoint video of human actors using multiple handheld kinects. *IEEE transactions on cybernetics*, 2013. 2, 3, 6
- [31] L. Zhang, J. Sturm, D. Cremers, and D. Lee. Real-time human motion tracking using multiple depth cameras. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 2389–2395. IEEE, 2012. 2, 6

Author	Context	Number Of RGB-D sensors in setup	Accuracy	Calibrated
Asteriadis <i>et al.</i> [3]	Motion Estimation	3	not specified	Yes
Berger <i>et al.</i> [6]	Motion Estimation	4	Reprojection Error Of $1.7px$	Yes
Fuhrmann <i>et al.</i> [10]	Motion Estimation	3	Deviation of $2 - 3cm$	Yes
Hossny <i>et al.</i> [11]	Motion Estimation	2	not specified	Yes (The authors provide a new autocalibration algorithm)
Santhanam <i>et al.</i> [21]	Motion Estimation	4	Deviation of $3mm$	Yes
Wilson <i>et al.</i> [29]	Motion Estimation	3	not specified	Yes
Ye <i>et al.</i> [30]	Motion Estimation	not specified	not specified	No
Zhang <i>et al.</i> [31]	Motion Estimation	2	Deviation of $20cm$	Yes
Alexiadis <i>et al.</i> [2]	Mesh Reconstruction	4	Reprojection Error Of $0.8px$	Yes
Berger <i>et al.</i> [5] and Berger <i>et al.</i> [4]	Mesh Reconstruction	3	not specified	Yes
Macknoja <i>et al.</i> [14]	Mesh Reconstruction	5	Deviation of $2.5cm$ at $3m$ distance	Yes
Lo <i>et al.</i> [13]	Mesh Reconstruction	2	not specified	Yes
Nakara <i>et al.</i> [16]	Mesh Reconstruction	2	Deviation of $3\%$ at $90^\circ$ spacing	No
Nakazawa <i>et al.</i> [17]	Mesh Reconstruction	4	not specified	Yes
Naveed [1]	Mesh Reconstruction	6	not specified	Yes
Olesen <i>et al.</i> [18]	Mesh Reconstruction	3	$60\%$ inlier at $8px$ Texlet spacing	Yes
Pancham <i>et al.</i> [19]	Mesh Reconstruction	2+	not specified	No
Rafibakhsh <i>et al.</i> [20]	Mesh Reconstruction	2	Deviation of $3.49cm$	Yes
Sumar <i>et al.</i> [26]	Mesh Reconstruction	2	Reprojection Error Of $5px$	No
Wang <i>et al.</i> [28]	Mesh Reconstruction	2	not specified	Yes
Caon <i>et al.</i> [8]	Recognition	3	not specified	Yes
Satta <i>et al.</i> [23]	Recognition	2	not specified	No
Satyavolu <i>et al.</i> [24]	Recognition	5	Deviation of $3cm$	Yes
Saputra <i>et al.</i> [22]	Recognition	2	Deviation of $10cm$	Yes
Susanto <i>et al.</i> [27]	Recognition	5	Deviation of $13cm$	Yes
Butler <i>et al.</i> [7]	Interference	2 and 3	Deviation of up to $3cm$	Yes
Faion <i>et al.</i> [9]	Interference	4	Deviation of $21mm$	Yes
Kainz <i>et al.</i> [12]	Interference	8	not specified	Yes
Maimone and Fuchs [15]	Interference	6	Deviation of $2mm$	No
Schroeder <i>et al.</i> [25] and Berger <i>et al.</i> [6]	Interference	4	Reprojection Error Of $1.7px$	Yes

Table 1. An Overview over different publications including multiple RGB-D sensors. 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.