A State of the Art Report on Kinect Sensor Setups in Computer Vision

Kai Berger[1], Stephan Meister[2], Rahul Nair[2], and Daniel Kondermann[2]

[1] OeRC Oxford, University of Oxford (firstname.lastname@oerc.ox.ac.uk)
[2] Heidelberg Collaboratory for Image Processing, University of Heidelberg (firstname.lastname@iwr.uni-heidelberg.de)

Abstract. During the last three years after the launch of the Microsoft Kinect® in the end-consumer market we have become witnesses of a small revolution in computer vision research towards the use of a standardized consumer-grade RGBD sensor for scene content retrieval. Beside classical localization and motion capturing tasks the Kinect has successfully been employed for the reconstruction of opaque and transparent objects. This report gives a comprehensive overview over the main publications using the Microsoft Kinect out of its original context as a decision-forest based motion-capturing tool.

1 Introduction

In early March 2010 Microsoft released a press text [54] that it would work together with PrimeSense, a Tel-Aviv based chip supplier, on a "groundbreaking optical-sensing and recognition technology to aid gesture control platforms." for the upcoming holidays. The goal of the project, internally known as "Project Natal" was to develop a new controller-free entertainment environment. Microsoft anticipated a paradigm shift on how people would interact with consumer-grade electronic devices.

The device itself was presented to a public audience at the E3 game convention. The device was launched in North America on November 4, 2010 and in Europe on November 10, 2010. By the beginning of 2012, 24 million units were sold. On February 1, 2012, Microsoft released the Kinect® for Windows SDK [53] and it is believed that more than 300 companies are working on apps that employ the Microsoft Kinect. In November 2010, Adafruit Industries funded an open-source driver development for Kinect. Although Microsoft initially disapproved their approach, they later clarified their position claiming that the USB connection was left open by design. Adafruit recognized Hèctor Martìn's work on a Linux driver that allows the use of both the RGB camera and depth sensitivity functions of the device. It is publicly available for download under the name *libfreenect* [62]. It is estimated that the OpenKinect community consists of roughly 2000 members who are contributing their time and code to the project. The code contributed to OpenKinect is made available under an Apache 2.0 or optional GPL2 license. Another open source API is provided via the OpenNI

framework of the OpenNI Organization [63] in which PrimeSense is a major contributor. In the middle of May 2013 Microsoft released a technical demo of the successor, Microsoft Kinect 2.0, which is based on Time-Of-Flight imaging. Both the availability of a consumer-grade RGBD sensor at a competitive price and the Open Source project that allowed to easily read out the essential streams from the sensor, quickly sparked an interest in the research community.

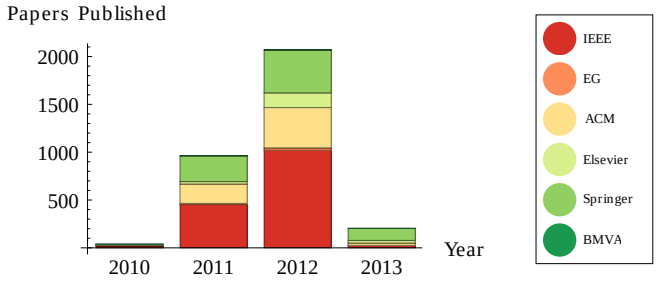

Fig. 1. The impact of the Microsoft Kinect in the computer vision field is significant: over the last three years, over 3000 papers related to the Microsoft Kinect have been published in renowned journals and proceedings (e.g., IEEE Explore, Digital Library of Eurographics, Proceedings of the ACM, Elsevier). Keywords associated with the Kinect include simultaneous localization and mapping, object reconstruction, multiple Kinect, interference mitigation, transparency and calibration.

Over the last three years a significant part of the published papers has been devoted to the use of the Kinect in a scientific context, Fig. 1. Over 3000 papers have been published in renowned journals and proceedings, e.g., Elsevier (208 papers), Eurographics (36 papers), ACM (651 papers), Springer (746 papers) or IEEE Explore (1518 papers), which publishes CVPR and ICCV proceedings among others. Of these, 276 papers refer to simultaneous localization and mapping problems and 227 are related to object reconstruction. Another 17 articles recognize the challenge that transparency, e.g. from a glass object, would pose on a sensor like the Kinect and proposed algorithms to reconstruct such transparent objects from depth streams from the Kinect. Finally, 47 papers address new ways to calibrate the Kinect. Further details about the deployment of a single Microsoft Kinect in academic context can be found in the manuscript submitted by Han et al. [27].

We recognize that there are still new ambitious research projects incorporating the Microsoft Kinect, e.g. the project Kinect@Home [2]. There, the user can help robotics and computer vision researchers around the world by scanning their office/living room environment with the Kinect. In return the user is delivered a 3D model of the very room.

The remainder of this state-of-the-art report is structured as follows: after reviewing the sensor itself in Section 2, we will introduce papers related to its use as a simultaneous localization and mapping tool in Section 3. Afterwards, we will

expand on motion capturing scenarios in which the Kinect has been employed, e.g. hand tracking, Section 4. Then, we will have a look into the research field that incorporates the Kinect as a tool to reconstruct non-opaque objects and motion, Section 5. In Section 6 we will present methods to improve or denoise Kinect depth maps while focusing on sensor fusion approaches. Finally, we will conclude and give an outlook, Section 7.

2 The Kinect 1.0 Sensor

Fig. 2. Typically, the Microsoft Kinect would be found in the living room of a Microsoft Xbox user. **Left:** typical usage scene, **Center:** infrared pattern, **Right:** colorcoded depth map.

The Microsoft Kinect is the first structured light sensor available for the consumer market. Designed as a motion sensing input device for the gaming console Microsoft XBox 360® the Kinect is intended to be used for gaming purposes. A typical usage environment can be seen in Figure 2(Left). With the Kinect it is possible for the XBox 360 to track movements of multiple players in a game. Its pattern emission technique was invented by PrimeSense and licensed by Microsoft for use in the Project Natal. The project OpenKinect provided the open source library libfreenect that enables PCs to use the Kinect as an input device via USB 2.0. This enabled users to experiment with an easy to access realtime capable depth tracking system. Compared to state-of-the-art depth capturing systems, e.g. time-of-flight (ToF) cameras, the system costs were negligible. With the success of the Kinect, other company's devices licensing the same technique from PrimeSense did appear. Asus introduced two devices called Xtion and Xtion LIVE with the underlying technique being the same as in the Kinect.

The coded light approach employed for the depth mapping is a simple and effective way to acquire depth data of a scene. A light, here an IR laser, projects a unique pattern onto the surface of the scene (see Figure 2(Center) for an example). This projection is recorded by a camera which is capable of capturing in the spectrum in which the pattern is emitted. Then, an integrated circuit computes the disparity for subpatterns by comparing them to their default positions at a given distance. For the disparity values the distance in meters for each pixel in the depth image can be computed. The structured light or active stereo approach is well known and has long been used by structured light scanners e.g. in the form of gray-codes for high precision depth measurements. The special pattern of the light used in the Kinect is particulary suited for fast disparity estimation using

block-matching and has been introduced by PrimeSense. In so far the Kinect suffers from the same depth estimation problems as other active or also passive stereo systems, mainly inaccurate depth at occlusion boundaries and problems with reflecting or transparent surfaces. A colorcoded representation of the depth values can be seen in Figure 2(Right).

3 SLAM and 3D Reconstruction

3D reconstruction and simultaneous localization and mapping (SLAM) are two closely connected fields of application which both can benefit from accurate depth data. Both can rely on either monoscopic reconstruction methods without prior depth information, sparse 3D data e.g. from laser rangefinders or dense depth maps e.g. from stereoscopic systems. Although systems utilizing only depth data or only visual data have been in use for decades, the integration of RGBD to make the systems more robust is a relatively new development. Apart from algorithms which were specifically designed for the Kinect we will also cover those that combine RGB and depth data in new ways and those which were inspired by these works even if they are not specifically limited to the Kinect.

A first step in both algorithm classes is the estimation of camera movement between consecutive frames. As shown by Handa et al. [28] tracking does generally benefit from high-frame rates alongside high resolution and low SNR. The Kinect sensor fills a niche in that it can supply dense depth maps in realtime. Examples for odometry algorithms which use depth data were presented by Kerl [41] or Steinbrücker [75]. Additionally, it has been shown by Newman and Ho [58] that visual features can effectively be used to solve the loop-closing problem in SLAM applications. The simultaneous availability of RGB and depth data can in this context be further exploited to calculate a dense scene flow [23]. Specific calibration considerations are discussed in [73] or [33]. Currently, there is no known SLAM system that uses multiple Kinects, although motion tracking with stationary cameras was demonstrated e.g. by Faion et al. [19] or Schönauer and Kaufmann [71].

One of the first methods to utilize the Kinect in a SLAM system is the framework presented by Henry et al. [30][31]. Here, features extracted from the RGB images are used for the initial camera pose estimation which is then refined by applying an iterative closest point algorithm (ICP) on the depth data. Hu et al. [34] use a similar approach but fall back to pure RGB based pose estimation if the depth features are insufficient, thereby adding the advantages of depth maps without inheriting their problems. Another approach was presented by Endres et al. [18] who also extract RGB features but then reproject these features into 3d to perform pose estimation in a closed form. All these algorithms can be used for online processing but unlike most recent developments which utilize GPU computation they are not real-time capable. Additionally, they do not always produce dense 3d representations like the following reconstruction algorithms as this is generally not necessary for localization tasks.

Accurate 3d reconstruction was until now a slow and expensive process as it was mostly based on laser or structured light scanners. The KinectFusion algorithm which was first introduced by Newcombe, Izadi et al. [57][38] and its subsequent improvements [66][32][83] represent a new direction in algorithm development as it is fast and depends only on commodity hardware. It creates an implicit voxel representation of a scene from the depth data using truncated signed distances functions. Each new view from the camera is registered using an Iterative Closest Point (ICP) algorithm. In that regard it behaves similar to other SLAM algorithms but the in-memory voxel representation allows for highly parallelized processing using GPUs. By providing a realtime 3D reconstruction method in the low to medium accuracy range (mm to cm regarding depth) it makes 3D scanning affordable for a wide field of potential users.

An analysis of the KinectFusion reconstruction performance has been performed by Meister et al. [52]. They compared the 3D meshes created by the KinectFusion system with high accuracy scans from LiDAR or structured light scanners to provide definite accuracy measures for mesh surfaces and derived values. The results suggest that the method is suitable even for applications where one would suspect an accuracy as high as possible to be mandatory. The geometric errors of 3D meshes created by KinectFusion can range from 10mm for small scenes (less than 1 m across, see Figure 3 for an example) to 80mm for room sized scenes. This may be too large for industrial inspection purposes but perfectly reasonable for the creation of synthetic test sequences for low-level image processing tasks, such as stereo matching or optical flow evaluation.

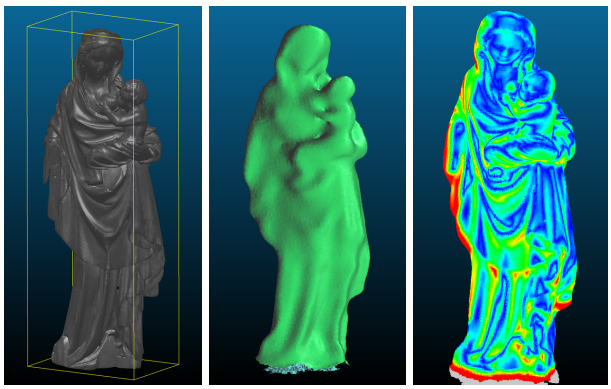

Fig. 3. Ground truth mesh, Kinect fusion mesh and euclidean surface error for scanned object from [52].

Despite it's impressive impact on both research and application alike the algorithm should not be considered a full SLAM solution. It's biggest drawbacks are the limited scan volume ($\approx 100 - 200m^3$ depending on graphics memory), the tendency to loose camera tracking in regions with few geometry features and the lack of explicit loop-closure handling. Some direct modifications of the algorithm try to alleviate these problems. Moving Volume Kinect by Roth et al. [66] allows the camera to leave the initial bounding volume but the basic limits for the 3d model still apply. Others like Kinfu Large Scale [32] or [87]

use more memory efficient data structures to represent the volume data, e.g. by using octrees. Kintinuous by Whelan et al. [83] continuously converts the volume data to point clouds for processing in main memory. This effectively removes any hard size limitations for the mapping volume. Whelan et al. also combined their system with the odometry estimation by Steinbrücker to make it more robust in case of missing geometric features [82]. This method is so far the only KinectFusion inspired algorithm that integrates RGB data. Bylow et al. [11] directly use the signed distance function of the voxel representation instead of ICP to estimate the camera movement more exactly. Keller et al. [40] drop the voxel representation altogether and use point-based fusion instead. Their approach handles the Kinect specific depth noise better and can handle dynamic scene content.

Other recent works try to combine SLAM with real-time capabilities and dense 3d reconstruction. Examples include the works by Lee et al. [44] who directly create a polygon representation from the acquired depth data or Henry et al. [29] who combine volumetric fusion with large-scale models. Finally, Stückler et al. [76] [77] use a different method based on a surfel representation of the environment. The camera pose estimation is also different in that it is estimated by a likelihood optimization approach on the surfel distribution. These recent developments suggest that the distinction between SLAM and 3D reconstruction may disappear in the near future as both algorithm types profit from improvements made to each other.

4 Motion Capturing Setups

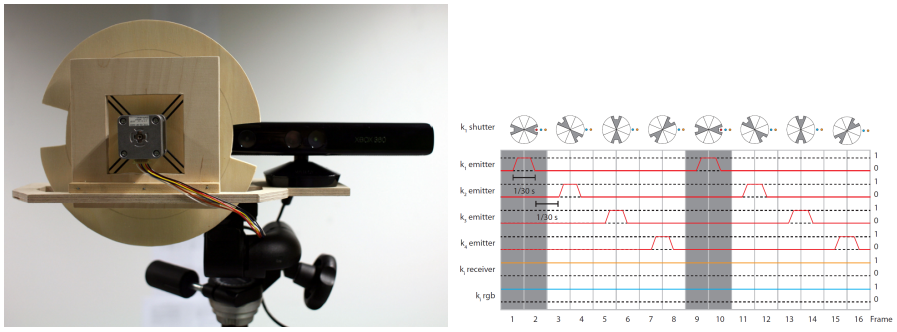

Fig. 4. An approach to incorporate multiple Kinects nondestructively in a motion capturing setup: An externally synced rolling shutter assigns one Kinect a unique time slot so that three other Kinects can capture as well. Such setups enable the capturing of obstructed motions or of motions with the actor not facing a camera. Red dots represent the emitters (projectors) while green dots represent receivers (cameras). Reproduced from [70].

Shotton et al. [72] introduced the Kinect and its underlying algorithm as a tool to capture the human pose from monocular depth images. Quickly thereafter, monocular motion capturing has gotten into the focus of the research community [22, 65, 60], with the Microsoft Kinect being the device to generate

datasets and benchmarks. What can be done with this research has been shown by Chen et al. [13]. Besides the tracking of limbs and joints quickly other research fields in monocular depth processing have emerged.

One interesting research direction for example is to use the Microsoft Kinect as a hand-tracking device. Oikonomidis [61] presents an approach based on particle swarms to discriminate between the palm and single fingers. Frati and his colleagues [21] assume the hand to always be closest to the camera and calculate convexity defects from the bounding box of the hand with the help of OpenCV while Reheja and his colleagues first detect the palm with a circular filter and then remove it to arrive at the shapes of individual fingertips in the depth image [64]. An interesting approach has been proposed by Van den Bergh et al. [6], who estimate the orientation of the hand from the orientation of the forearm in the depth image. The posture itself is estimated by employing an Average Neighborhood Margin Maximization (ANMM) algorithm [80].

With the Microsoft Kinect it is also possible to capture facial movements. Zollhofer et al. [89] showed how to fit deformable facial meshes to depth data captured from human faces by relying on feature points (eyes, nose) in the depth data. Leyvand et al. also examine the face recognition of identical twins given depth and motion data from the Microsoft Kinect [46].

In 2011, Berger and his colleagues showed, that it is also possible to employ multiple Microsoft Kinects in one scene for motion capturing research [5]. Their incentive was to enable the capturing of partially obstructed poses, e.g. from persons facing away from the camera or in small rooms. Using a specifically tailored external hardware shutter [70] they were able to reduce the sensor noise introduced from neighboring Kinects, Fig. 4. Their approach relied on synchronized rolling shutters for up to four devices. This idea was quickly adopted and further developed by Maimone and Fuchs [50] in a shake and sense approach: each Kinect sensor would slightly rotate around its up vector introducing scene motion to the imaged scene except for its own projected pattern which always moves accordingly. Thus, the accuracy of the depth image generated from its own pattern would increase due to blurred out sensor noise from other Kinects. The motion would be accounted for from the Kinect's inertial sensor data. This approach was further refined by Butler and his colleagues [10] who basically hot-melt glued a motor to each device to introduce arbitrary motion.

5 Opaque and Transparent Reconstruction

With the availability of accurate depth data, the complete 3D reconstruction of objects with the consumer-grade Kinect became a popular research branch. For example, Tam and his colleagues [78] register point clouds captured with the Kinect to each other.

However, the reconstruction need not necessarily be restricted to opaque objects. Lysenkov and his colleagues [48] describe an approach to recognize transparent objects, e.g. a water glass, and to recognize its pose from the input images of a Kinect device. Due to reflection and transmission the IR pattern shone onto

the transparent objects is not usable for depth estimation. Consequently, pixel regions of the projected object in the depth image obtain invalid values, e.g. appear black. They use a key observation: Transparent and opaque objects create surface and silhouette edges. Image edgels corresponding to a silhouette edge can be detected at the boundary between the invalid and valid depth pixels. To recognize transparent objects one can reconstruct it by moving the Kinect 360° around the object or by comparing it to a similar mesh in a database. They, however, decide to register it beforehand by powdering it and thus making it temporarily opaque. The silhouettes of the registered object are then used for training. During the test phase later, they compare the silhouette edges created by invalid pixels in the depth images with the silhouettes in the database using Procrustes Analysis as proposed by [51]. When a non-powdering approach is pursued, the authors stress that it is important to provide additional calibration information [47] for the Kinect in order to reconstruct its location to the transparent object, whose only viable information are the silhouette edges retrieved from the depth images. Another approach to reconstruct transparent objects with the Kinect is to incorporate the RGB-sensor. Chiu et al. [15] propose to calibrate the RGB-camera with the IR-camera to arrive at a multi-modal stereo image (i.e., depth, and the stereo from disparity between the RGB- and IR-camera).

When the object to be reconstructed becomes time-varying, it is impossible to powder and capture it beforehand. In their work, Berger et al. [4] examined the possibilities to reconstruct transparent gas flows using the Kinect. They ruled out seeding particles and decided to follow a Background-oriented Schlieren approach. The projected IR-pattern of each Kinect is hereby used as the background pattern. The silhouette boundaries would become visible in the depth sensor by the index gradient between the flowing gas, there propane, and its surrounding medium (air). As propane obtains a refractive index of roughly 1.34 the difference to the surrounding air would be sufficiently high enough to introduce noticeable pixel deviations at a distance of $3m$ between scene walls and the Kinect camera. They concluded, that, when they would place three Kinects in an half-arc around the flowing gas and projection walls at a fixed distance opposite to it, they could detect difference in the depth images that would suffice for silhouettes. Using the silhouettes of each Kinect they could enclose the gas volume in the reconstructed visual hull for each frame. The silhouette generation relied on fitting polynomials from left and right in each image [1, 4]. In further research they concluded that it is also viable to directly use the deviations in the IR-images for the silhouette reconstruction, by relying on a sparse spot-based optical flow algorithm [69].

6 Enhancing Depth data

Although the Kinect delivers RGBD data of a sufficient quality for many applications, it is far from perfect. For example, as the projector is located to the right of the cameras, no depth data can be obtained in areas to the left of occlusion

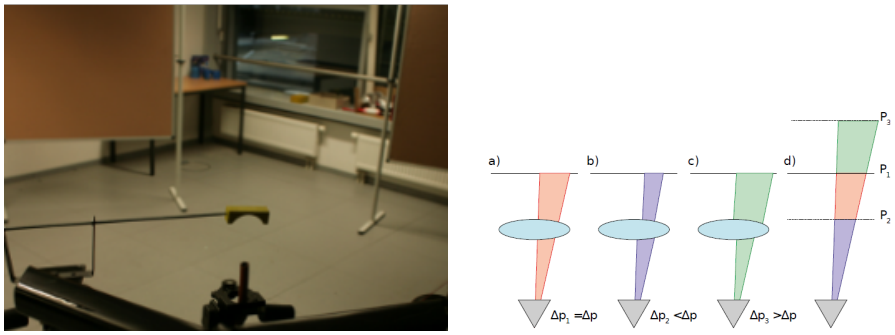

Fig. 5. The reconstruction of non-opaque motion. Three Kinects are placed in a circular half-arc around propane gas flow, projection walls opposite to each Kinect. As the Kinects do not interfere destructively with each other, meaningful information can be retrieved for each sensor. The refractive index gradient present in the scene would result in detectable depth deviations in each Kinect's depth image stream. Reproduced from [4]

boundaries due to shadowing. If the depth map is then additionally registered to the RGB image, further information is lost. Other effects which are present throughout the image are errors due to the sparsity of the point pattern, the block size used for matching and the unknown smoothing that may additionally be applied to the raw data. Most of these errors can best be observed at depth edges. They lead to inaccurate depth boundaries, blobbing artifacts and a reduced effective lateral resolution. Also like every other active depth imaging technique the Kinect relies on the reflected light being of sufficient intensity. This is not the case with dark IR absorbing surfaces that may additionally lie at an angle to the camera or when strong IR light sources such as direct sun light are present in the scene [55].

The question remains whether there is a real need for better quality or higher resolution depth data. ICP [7][86] which is at the core of many pose and 3d reconstruction algorithms using Kinect, will produce better results given better input data. Also, accurate silhouette information is a strong cue used for 3d reconstruction [43].Some applications even depend on good initial depth data. As an example the visual effects industry frequently requires dynamic scene geometry at resolutions ranging from Full HD to 4K [39]. Current depth cameras meet the dynamic imaging requirement but fail to provide the necessary lateral resolution. In the following we will review the various lines of research dealing with the enhancement of depth images. Often, the papers presented deal with Time of Flight data instead of Kinect. Many of these algorithms work on the depth images and thus can be directly applied to Kinect data. Others also take into account the noise characteristics of Time of Flight sensors which are generally quite different from those of the Kinect. Here, the noise model used must be replaced with the Kinect noise model such es the empirical model recently presented by Nguyen et al. [59].

Depth data denoising as a subdiscipline of image denoising has progressed significantly and many edge preserving denoising techniques can be applied directly to range images. Examples would be diffusion based filters [81], non local means [9] or bilateral filtering [79]. Unlike RGB images, depth images are generally considered to be comparitively smooth with few distinct edges [35][84]. This property allows for a much stronger regularization than would be possible in RGB images. Lenzen et al. [45] apply an adaptive first and second order total variation approach to regularize depth data while retaining edges and slopes. Schoner et al. [68] apply a clustering approach to identify regions with similar properties. Aodha et al. [49] learn the relation between noisy input images and filtered output using decision tree ensembles [8].

As mentioned above, Kinect depth data contains many invalid pixels. To alleviate this problem, hole filling strategies which are related to image inpainting can be employed. Danciu et al. presented a single-frame method based on morphological filters [17]. Other Methods additionally use temporal information to make the inpainting more robust. Xu et al. first detect moving objects to improve edge stability before filling in holes [85], while Camplani and Salgado use bilateral filtering in combination with a temporal consistency constraint [12].

A different method to enhance Kinect data is to apply a sensor fusion approach by adding additional depth imaging modalities to create superresolution depth images. The sensor fusion methods can be differentiated by the employed camera setup. As strategies for using multiple Kinects have been discussed in Section 5 we will therefore limit ourselves to approaches using one or two additional RGB cameras. As the Kinect sensor itself includes a RGB camera and an IR camera, it can be used directly for RGBD fusion. Often though, an external RGB camera with a higher resolution is used for the fusion approach. After aligning the RGB and IR camera employing standard camera calibration techniques the main assumption is that depth edges often coincide with RGB edges. Chen et al. [14] for example employ cross bilateral filtering to smooth the resulting depth maps. Huhle et al. propose a graphical model with data terms based on RGB and depth gradient strength in [36] and in [37] adapted non local means filtering to encompass the additional data. Chiu et al. [15] on the other hand use the cross modal stereo information between the IR and the RGB sensor directly.

Most works which combine depth cameras with a regular passive stereo setup have been done with ToF imagers but as already mentioned the methods can be adapted to Kinect most of the time. One exception it the recently presented method by Somanath et al. [74] which uses a kinect to improve stereo depth estimates in ambiguous or low-textured regions. These methods use the range imaging data to initialize stereo matching and impose constraints on the search range depending on the depth budget and stereo noise model. Local methods [42],[24],[3],[26],[16],[56] combine the stereo and the range imaging data term on a per pixel level. Gudmundsson et al.[24] apply a hierarchical stereo matching algorithm directly on the remapped depth data without considering uncertainties. Kuhnert et al.[42] and Hahne et al.[26] compute binary confidences in the

depth image and let stereo refine the result in regions with low confidence. Nair et al.[56] and Dal Mutto et al. [16] locally combine confidences from both stereo and the depth image into the the stereo matching framework. Global methods [20],[56],[88],[25] additionally apply spatial regularization techniques to propagate more information to regions with low stereo or depth image confidence. Inference of the global energy is then done using different optimization methods such as graph cuts[25], semi global optimization[20], MAP-MRF [88] or by minimizing the total variation[56],[67].

7 Conclusion

This state of the art report has reviewed the Kinect as a consumer-grade motion capturing toolkit and recognized its impact in the computer vision community. The output of the Kinect, depth-, RGB- and IR-images at realtime framerate enabled researchers to use the device in various scenarios. Simultaneous localization and mapping (SLAM) in the context of robotics and object reconstruction showed that the Kinect sensor fills a niche in that it can supply dense depth maps in realtime. Out of its intended context the Kinect was employed to track gestures and recognize faces. In small room environments it was shown that multiple Kinect sensors could capture motion without interfering destructively with each other thus enabling the capturing of obstructed motions or the motions of actors facing away from one camera. Recently, it was examined if non-opaque objects can be reconstructed as well. By relying on silhouette edges present in the depth images, e.g. around invalid depth pixel, the question could be answered positively for glass objects and gas flows. We conclude that this capturing has made an impact to the community that is unprecedented and sparked very creative research ideas. Additionally many advancements in the field of sensor fusion or depth map denoising e.g. from time-of-flight imaging can be applied to the Kinect camera to improve its accuracy.

Although now, 3 years later, a new generation of consumer-grade motion capturing devices is ready to be deployed and to challenge the position of the Microsoft Kinect. We believe that the impact of the Kinect and similar devices will continue to increase in the next years and that it will become the standard prototyping-research tool on every desktop in the vision community.

8 Acknowledgements

This work has been partially funded by the Intel Visual Computing Institute, Saarbrücken (IVCI) as part of the Project "Algorithms for Low Cost Depth Imaging" and the Engineering and Physical Sciences Research Council (EPSRC) grant on Video Visualization. Kinect and Xbox 360 are registered trademarks of Microsoft Corporation. This is an independent publication and is not affiliated with, nor has it been authorized, sponsored, or otherwise approved by Microsoft Corporation.

References

1. Albers, M., Berger, B.K., Magnor, E.P.D.I.M.: The capturing of turbulent gas flows using multiple kinects. Bachelor thesis, Technical University Braunschweig (2012)
2. Aydemir, A., Henell, D., Jensfelt, P., Shilkrot, R.: Kinect@ home: Crowdsourcing a large 3d dataset of real environments. In: 2012 AAAI Spring Symposium Series (2012)
3. Bartczak, B., Koch, R.: Dense depth maps from low resolution time-of-flight depth and high resolution color views. Advances in Visual Computing pp. 228–239 (2009)
4. Berger, K., Ruhl, K., Albers, M., Schroder, Y., Scholz, A., Kokemuller, J., Guthe, S., Magnor, M.: The capturing of turbulent gas flows using multiple kinects. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 1108–1113. IEEE (2011)
5. Berger, K., Ruhl, K., Brümmer, C., Schröder, Y., Scholz, A., Magnor, M.: Markerless motion capture using multiple color-depth sensors. In: Proc. Vision, Modeling and Visualization (VMV). vol. 2011, p. 3 (2011)
6. Van den Bergh, M., Carton, D., De Nijs, R., Mitsou, N., Landsiedel, C., Kuehnlenz, K., Wollherr, D., Van Gool, L., Buss, M.: Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In: RO-MAN, 2011 IEEE. pp. 357–362. IEEE (2011)
7. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Transactions on pattern analysis and machine intelligence 14(2), 239–256 (1992)
8. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
9. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 2, pp. 60–65. IEEE (2005)
10. Butler, D.A., Izadi, S., Hilliges, O., Molyneaux, D., Hodges, S., Kim, D.: 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. pp. 1933–1936. ACM (2012)
11. Bylow, E., Sturm, J., Kerl, C., Kahl, F., Cremers, D.: Real-time camera tracking and 3d reconstruction using signed distance functions. In: Robotics: Science and Systems Conference (RSS) (June 2013)
12. Camplani, M., Salgado, L.: Efficient spatio-temporal hole filling strategy for kinect depth maps. In: IS&T/SPIE Electronic Imaging. pp. 82900E–82900E. International Society for Optics and Photonics (2012)
13. Chen, J., Izadi, S., Fitzgibbon, A.: Kinêtre: animating the world with the human body. In: Proceedings of the 25th annual ACM symposium on User interface software and technology. pp. 435–444. ACM (2012)
14. Chen, L., Lin, H., Li, S.: Depth image enhancement for kinect using region growing and bilateral filter. In: Pattern Recognition (ICPR), 2012 21st International Conference on. pp. 3070–3073. IEEE (2012)
15. Chiu, W.C., Blanke, U., Fritz, M.: Improving the kinect by cross-modal stereo. In: 22nd British Machine Vision Conference (BMVC) (2011)
16. Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.M.: A probabilistic approach to tof and stereo data fusion. In: 3DPVT. Paris, France (May 2010)
17. Danciu, G., Banu, S.M., Caliman, A.: Shadow removal in depth images morphology-based for kinect cameras. In: System Theory, Control and Computing (ICSTCC), 2012 16th International Conference on. pp. 1–6. IEEE (2012)

18. Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., Burgard, W.: An evaluation of the rgb-d slam system. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on. pp. 1691–1696. IEEE (2012)
19. Faion, F., Friedberger, S., Zea, Antonio Hanebeck, U.D.: Intelligent sensor-scheduling for multi-kinect-tracking. In: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS) (2012)
20. Fischer, J., Arbeiter, G., Verl, A.: Combination of time-of-flight depth and stereo using semiglobal optimization. In: Int. Conf. on Robotics and Automation (ICRA). pp. 3548–3553. IEEE (2011)
21. Frati, V., Prattichizzo, D.: Using kinect for hand tracking and rendering in wearable haptics. In: World Haptics Conference (WHC), 2011 IEEE. pp. 317–321. IEEE (2011)
22. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 415–422. IEEE (2011)
23. Gottfried, J.M., Fehr, J., Garbe, C.: Computing range flow from multi-modal kinect data. Advances in Visual Computing pp. 758–767 (2011)
24. Gudmundsson, S., Aanaes, H., Larsen, R.: Fusion of stereo vision and time-of-flight imaging for improved 3d estimation. IJISTA 5(3), 425–433 (2008)
25. Hahne, U., Alexa, M.: Combining time-of-flight depth and stereo images without accurate extrinsic calibration. IJISTA 5(3), 325–333 (2008)
26. Hahne, U., Alexa, M.: Depth imaging by combining time-of-flight and on-demand stereo. Dynamic 3D Imaging pp. 70–83 (2009)
27. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: A review. In: Transactions on Cybernetics. IEEE (2013)
28. Handa, A., Richard, N.A., Angeli, A., Davison, A.J.: Real-Time camera tracking: when is high frame-rate best? In: 12th European Conference on Computer Vision (ECCV) (2012), http://link.springer.com/chapter/10.1007/978-3-642-33786-4_17
29. Henry, P., Fox, D., Bhowmik, A., Mongia, R.: Patch Volumes: Segmentation-based Consistens Mapping with RGB-D Cameras. In: 3D Vision 2013 (3DV), International Conference on (2013)
30. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In: the 12th International Symposium on Experimental Robotics (ISER). vol. 20, pp. 22–25 (2010)
31. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. The International Journal of Robotics Research 31(5), 647–663 (2012)
32. Heredia, F., Favier, R.: Point cloud library developers blog, kinfu large scale(june 18, 2012). http://www.pointclouds.org/blog/srcs/
33. Herrera C, D., Kannala, J., Heikkilä, J.: Accurate and practical calibration of a depth and color camera pair. In: Computer Analysis of Images and Patterns. pp. 437–445. Springer (2011)
34. Hu, G., Huang, S., Zhao, L., Alempijevic, A., Dissanayake, G.: A robust rgb-d slam algorithm. In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on (2012)
35. Huang, J., Lee, A.B., Mumford, D.: Statistics of range images. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. vol. 1, pp. 324–331. IEEE (2000)
36. Huhle, B., Fleck, S., Schilling, A.: Integrating 3d time-of-flight camera data and high resolution images for 3dtv applications. In: Proc. 3DTV Conf. IEEE (2007)

37. Huhle, B., Schairer, T., Jenke, P., Straßer, W.: Robust non-local denoising of colored depth data. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on. pp. 1–7. IEEE (2008)
38. Izadi, S., Newcombe, R.A., Kim, D., Hilliges, O., Molyneaux, D., Hodges, S., Kohli, P., Shotton, J., Davison, A.J., Fitzgibbon, A.: KinectFusion: real-time dynamic 3D surface reconstruction and interaction. In: ACM SIGGRAPH 2011 Talks. p. 23. ACM (2011)
39. Kate Solomon - techradar.com: Meerkats to go Ultra HD in BBC's first 4K broadcast, http://www.techradar.com/news/tv/television/meerkats-to-go-ultra-hd-in-bbcs-first-4k-broadcast-1127915/
40. Keller, M., Lefloch, D., Lambers, M., Izadi, S., Weyrich, T., Kolb, A.: Real-time 3D Reconstruction in Dynamic Scenes using Point-based Fusion. In: 3D Vision 2013 (3DV), International Conference on (2013)
41. Kerl, C., Sturm, J., Cremers, D.: Robust odometry estimation for rgb-d cameras. In: Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA) (May 2013)
42. Kuhnert, K., Stommel, M.: Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction. In: Int. Conf. on Intelligent Robots and Systems. pp. 4780–4785. IEEE (2006)
43. Laurentini, A.: The visual hull concept for silhouette-based image understanding. Pattern Analysis and Machine Intelligence, IEEE Transactions on 16(2), 150–162 (1994)
44. Lee, T., Lim, S., Lee, S., An, S., Oh, S.: Indoor mapping using planes extracted from noisy rgb-d sensors. In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on (2012)
45. Lenzen, F., Schaefer, H., Garbe, C.: Denoising time-of-flight data with adaptive total variation. In: Advances in Visual Computing. LNCS, vol. 6938, pp. 337–346. Springer (2011)
46. Leyvand, T., Meekhof, C., Wei, Y.C., Sun, J., Guo, B.: Kinect identity: Technology and experience. Computer 44(4), 94–96 (2011)
47. Lysenkov, I., Eruhimov, V.: Pose refinement of transparent rigid objects with a stereo camera. In: 22th International Conference on Computer Graphics and Vision, GraphiCon'2012 (2012)
48. Lysenkov, I., Eruhimov, V., Bradski, G.: Recognition and pose estimation of rigid transparent objects with a kinect sensor. In: Robotics: Science and Systems VIII, Sydney, Australia (2012)
49. Mac Aodha, O., Campbell, N.D., Nair, A., Brostow, G.J.: Patch based synthesis for single depth image super-resolution. In: 12th European Conference on Computer Vision. ECCV (2012)
50. Maimone, A., Fuchs, H.: Reducing interference between multiple structured light depth sensors using motion. In: Virtual Reality Workshops (VR), 2012 IEEE. pp. 51–54. IEEE (2012)
51. Mardia, K., Dryden, I.: The statistical analysis of shape data. Biometrika 76(2), 271–281 (1989)
52. Meister, S., Izadi, S., Kohli, P., Hämmerle, M., Rother, C., Kondermann, D.: When can we use kinectfusion for ground truth acquisition? In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, Workshops & Tutorials (2012)
53. Microsoft Corporation: Kinect for windows sdk, http://www.microsoft.com/en-us/kinectforwindows/

54. Microsoft News Center: Microsoft press release (mar 2010), http://www.microsoft.com/en-us/news/press/2010/mar10/03-31PrimeSensePR.aspx
55. Microsoft Xbox support: Room lighting conditions for kinect, http://support.xbox.com/en-US/xbox-360/kinect/lighting/
56. Nair, R., Lenzen, F., Meister, S., Schäfer, H., Garbe, C., Kondermann, D.: High accuracy tof and stereo sensor fusion at interactive rates. In: Computer Vision–ECCV 2012. Workshops and Demonstrations. pp. 1–11. Springer (2012)
57. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality. vol. 7, pp. 127–136 (2011)
58. Newman, P., Ho, K.: Slam-loop closing with visually salient features. In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. pp. 635–642. IEEE (2005)
59. Nguyen, C.V., Izadi, S., Lovell, D.: Modeling kinect sensor noise for improved 3d reconstruction and tracking. In: 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on. pp. 524–530. IEEE (2012)
60. Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., Kohli, P.: Decision tree fields. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 1668–1675. IEEE (2011)
61. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. BMVC, Aug 2 (2011)
62. Openkinect Project: libfreenect, http://openkinect.org/
63. OpenNI: Openni framework, http://www.openni.org
64. Raheja, J.L., Chaudhary, A., Singal, K.: Tracking of fingertips and centers of palm using kinect. In: Computational Intelligence, Modelling and Simulation (CIMSiM), 2011 Third International Conference on. pp. 248–252. IEEE (2011)
65. Raptis, M., Kirovski, D., Hoppe, H.: Real-time classification of dance gestures from skeleton animation. In: Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation. pp. 147–156. ACM (2011)
66. Roth, H., Vona, M.: Moving volume kinectfusion. In: British Machine Vision Conf.(BMVC),(Surrey, UK) (2012)
67. Ruhl, K., Klose, F., Lipski, C., Magnor, M.: Integrating approximate depth data into dense image correspondence estimation. In: Proceedings of the 9th European Conference on Visual Media Production. pp. 26–31. ACM (2012)
68. Schoner, H., Moser, B., Dorrington, A.A., Payne, A.D., Cree, M.J., Heise, B., Bauer, F.: A clustering based denoising technique for range images of time of flight cameras. In: Computational Intelligence for Modelling Control & Automation, 2008 International Conference on. pp. 999–1004. IEEE (2008)
69. Schröder, Y., Berger, K., Magnor, M.: Super resolution for active light sensor enhancement. Bachelor thesis, University of Braunschweig (Mar 2012)
70. Schröder, Y., Scholz, A., Berger, K., Ruhl, K., Guthe, S., Magnor, M.: Multiple kinect studies. Computer Graphics (2011)
71. Schnauer, C., Kaufmann, H.: Wide area motion tracking using consumer hardware. In: Proceedings of Workshop on Whole Body Interaction in Games and Entertainment, Advances in Computer Entertainment Technology (ACE 2011), Lisbon, Portugal. (2011)

72. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 1297–1304. IEEE (2011)
73. Smisek, J., Jancosek, M., Pajdla, T.: 3d with kinect. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 1154–1160. IEEE (2011)
74. Somanath, G., Cohen, S., Price, B., Kambhamettu, C.: Stereo+Kinect for High Resolution Stereo Correspondences. In: 3D Vision 2013 (3DV), International Conference on (2013)
75. Steinbrücker, F., Sturm, J., Cremers, D.: Real-time visual odometry from dense rgb-d images. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 719–722. IEEE (2011)
76. Stuckler, J., Behnke, S.: Integrating depth and color cues for dense multi-resolution scene mapping using rgb-d cameras. In: Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on. pp. 162–167. IEEE (2012)
77. Stückler, J., Behnke, S.: Multi-resolution surfel maps for efficient dense 3d modeling and tracking. Journal of Visual Communication and Image Representation (2013)
78. Tam, G., Cheng, Z.Q., Lai, Y.K., Langbein, F., Liu, Y., Marshall, A., Martin, R., Sun, X.F., Rosin, P.: Registration of 3d point clouds and meshes: A survey from rigid to non-rigid. Visualization and Computer Graphics, IEEE Transactions on PP, Issue 99, 1 (2012)
79. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Computer Vision, 1998. Sixth International Conference on. pp. 839–846. IEEE (1998)
80. Wang, F., Zhang, C.: Feature extraction by maximizing the average neighborhood margin. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. pp. 1–8. IEEE (2007)
81. Weickert, J.: Anisotropic diffusion in image processing, vol. 1. Teubner Stuttgart (1998)
82. Whelan, T., Johannsson, H., Kaess, M., Leonard, J.J., McDonald, J.: Robust real-time visual odometry for dense rgb-d mapping. In: IEEE Intl. Conf. on Robotics and Automation, ICRA, Karlsruhe, Germany (2013)
83. Whelan, T., Kaess, M., Fallon, M., Johannsson, H., Leonard, J., McDonald, J.: Kintinuous: Spatially extended kinectfusion. Tech. Rep. MIT-CSAIL-TR-2012-020, CSAIL Technical Reports (2012), http://hdl.handle.net/1721.1/71756
84. Woodford, O., Torr, P., Reid, I., Fitzgibbon, A.: Global stereo reconstruction under second-order smoothness priors. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(12), 2115–2128 (2009)
85. Xu, K., Zhou, J., Wang, Z.: A method of hole-filling for the depth map generated by kinect with moving objects detection. In: Broadband Multimedia Systems and Broadcasting (BMSB), 2012 IEEE International Symposium on. pp. 1–5. IEEE (2012)
86. Yang, C., Medioni, G.: Object modelling by registration of multiple range images. Image and Vision Computing 10(3), 145–155 (1992)
87. Zeng, M., Zhao, F., Zheng, J., Liu, X.: A memory-efficient kinectfusion using octree. In: Computational Visual Media, pp. 234–241. Springer (2012)
88. Zhu, J., Wang, L., Yang, R., Davis, J., et al.: Reliability fusion of time-of-flight depth and stereo for high quality depth maps. TPAMI (99), 1–1 (2011)
89. Zollhöfer, M., Martinek, M., Greiner, G., Stamminger, M., Süßmuth, J.: Automatic reconstruction of personalized avatars from 3d face scans. Computer Animation and Virtual Worlds 22(2-3), 195–202 (2011)