Low-cost commodity depth sensor comparison and accuracy analysis

Timo Breuer, Christoph Bodensteiner, Michael Arens

Fraunhofer IOSB, Gutleuthausstrasse 1, Ettlingen, Germany

ABSTRACT

Low cost depth sensors have been a huge success in the field of computer vision and robotics, providing depth images even in untextured environments. The same characteristic applies to the Kinect V2, a time-of-flight camera with high lateral resolution. In order to assess advantages of the new sensor over its predecessor for standard applications, we provide an analysis of measurement noise, accuracy and other error sources with the Kinect V2. We examined the raw sensor data by using an open source driver. Further insights on the sensor design and examples of processing techniques are given to completely exploit the unrestricted access to the device.

Keywords: 2.5D time-of-flight continuous modulation Kinect2

1. INTRODUCTION

The original Kinect made depth imaging available to a broad public. Its structured light sensor provides robust depth measurements even in textureless or poorly lit scenes, making it more reliable than stereo camera setups. Due to mass production the device is highly affordable, with prices being about one order of magnitude lower, compared to competing devices. Two of the most famous applications are human pose estimation[1], which has been the original use case, as a controllerless user interface for the XBox 360 and KinectFusion[2], an application that tracks the device's pose with 6 degrees of freedom and performs a dense 3d reconstruction of an unknown environment. Microsoft strives to continue this success with the release of the Kinect V2. Both devices utilize active illumination, which is captured by a camera specially suited for the light source's spectrum. They also feature a regular RGB camera in order to colorize the depth image by coordinate mapping. As opposed to the original sensor, a different measuring principle called continuous modulation, provides the time of flight of a light impulse for every single pixel.

This paper is organized as follows. In Section 2 we give an overview of related work covering the Kinect V1 as well as different ToF cameras. There is no literature available on the Kinect V2 as it has not been available to the market at the time of writing. Section 3 introduces the continuous modulation principle, applied by the Kinect V2. Some of the specifications are given and compared to other depth cameras. In Section 4 we analyze the main characteristics, such as resolution, accuracy and precision by using a calibration target and model fitting tests. When feasible, the same tests are performed on a Kinect V1-like device. This allows a direct comparison of both sensor generations and shows which devices is more suitable in different scenarios. We used an open source driver to record depth measurements from the device, which also provided raw sensor measurements, currently not exposed by the official SDK. Yet unpublished specifications and examples of preprocessing techniques are given, that could not have been obtained without the unprocessed data.

Further author information: (Send correspondence to T.B.)
T.B.: E-mail: timo.breuer@iosb.fraunhofer.de, Telephone: +49 7243 992-185

Electro-Optical Remote Sensing, Photonic Technologies, and Applications VIII; and Military Applications in Hyperspectral
Imaging and High Spatial Resolution Sensing II, edited by Gary Kamerman, Gary J. Bishop, et al., Proc. of SPIE
Vol. 9250, 92500G · © 2014 SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2067155

2. PREVIOUS WORK

Khoshelham[3] performed an analysis of the Kinect V1. In order to obtain intrinsic camera parameters, a calibration target was recorded at different orientations by the IR camera while the pattern projector was obstructed. A point cloud was generated by reprojecting every pixel of the depth image into 3D space. A high-end laser scanner was used to capture precise groundtruth data, which was registered against the point cloud using the iterative closest points method. In addition a planar object was recorded at varying distances and a plane was fitted to the data. In both cases the residuals of registration and model fitting approaches showed to be proportional to the squared distance within the operation range of the sensor. Martinez and Stiefelhagen[4] dissected the architecture of the depth estimation algorithm by extracting the projected dot pattern, acquisition of the raw IR image and doing stereo matching in software. Different approaches and parameters were tested, yielding a strong cue, that a filtered version of the 1280x1024 pixel camera image is used to perform block matching against the reference dot pattern, because, in this manner, the standard depth output of the sensor could be reproduced with high accuracy. They suggest, that the dot pattern is to sparse to provide accurate depth measurements for every pixel and that a different stereo matching algorithm (for example Semi-Global Matching[5]) would enable a higher depth resolution, due to better subpixel precision. While the Kinect V2 is a new device and is being made publicly available to researchers in summer 2014, continuous modulation is already used in many TOF cameras. Lefloch et al.[6] provide an excellent overview with a full mathematical description of the measuring principle, as well as possible error sources, such as motion artifacts or multiple returns. They also give a short survey of calibration and post processing approaches. A separate lateral and depth calibration of a TOF camera was performed by Lindner and Kolb[7] and by Kahlmann et al.[8] with planar calibration targets. Lindner[7] has used a chessboard pattern for lateral calibration, but the detection performance is strongly limited by the cameras resolution of only 160x120 pixels. Depth calibration was done by using a low resolution TOF camera as a reference device. Kahlmann[8] used an active calibration target featuring near infrared LEDs, thus enabling high precision subpixel detection and precise lateral calibration. Unfortunately, some TOF cameras filter out illumination not correlated to its modulation frequency, making a direct precise localization of the LEDs impossible. A track line with a trolley, localized with sub-millimeter precision by an interferometer, was used to move a target in order to acquire groundtruth data for depth calibration. A joint lateral and depth calibration was done by Schiller et al.[9] A planar calibration target was captured by the TOF and several standard CCD cameras, whose intrinsic and extrinsic camera parameters are initially guessed by standard camera calibration. Afterwards synthetic intensity and depth images of the calibration target were used to optimize all parameters. The model for depth synthesis was further refined by Lindner[10] incorporating different error sources in order to enhance the calibration and depth accuracy.

3. SENSOR

3.1 Continuous modulation principle

Lefloch[6] describes TOF cameras as a combination of a light source emitting a modulated signal and a special pixel array. These smart pixels are able to directly correlate the incident light signal $r(t)$ with the reference signal $s(t)$.

$$c(\tau) = r \otimes s = \lim_{T \to \infty} \int_{-T/2}^{T/2} r(s) \times s(t + \tau)\, dt \tag{1}$$

where τ is a phase offset used to sample the correlation function at different positions. Usually a sinusoidal reference signal is used,

$$s(t) = \ cos(\omega t), r(t) = I + A\cos(\omega t - \phi) \tag{2}$$

where ω is the modulation frequency, A is the amplitude of the captured component of the illumination correlated to the reference signal, I is the uncorrelated illumination (for example scene illumination not originating from the device itself) and ϕ is the phase offset of the returning signal due to the round-trip time of the light.

Four samples of the correlation function with $\tau_i = i\pi/2\omega$ yield two correlation values with suppressed background illumination, which can be used to calculate ϕ and a.

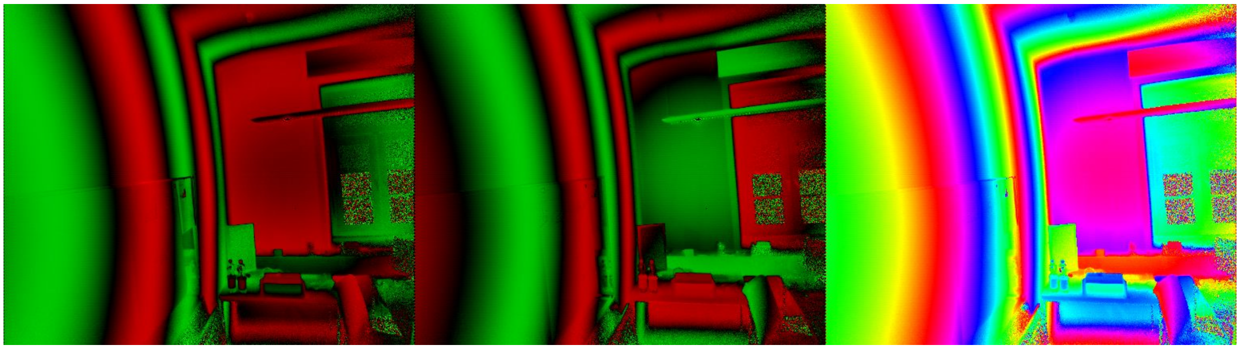

Figure 1. Visualization of the correlation coefficients and the corresponding phase for a modulation frequency of 120 MHz. Positive correlation coefficients are mapped to the green channel, negative coefficients to the red channel. The phase is mapped to the hue channel of an HSV image with full saturation and intensity. The three images correspond to the first argument, second argument and function value of atan2 in Equation 5. The scene is shown again as an IR image in Figure 5.

$$\begin{aligned} \phi &= atan2(c(\tau_3) - c(\tau_1), c(\tau_0) - c(\tau_2) + \pi \\ a &= \frac{1}{2}\sqrt{(c(\tau_3) - c(\tau_1))^2 + (c(\tau_0) - c(\tau_2)^2} \end{aligned} \tag{3}$$

The distance can calculated by

$$d = \frac{\lambda_m}{4\pi}\phi, \tag{4}$$

with λ_m being the modulation wavelength. A lower wavelength will lead to more accurate results in the presence of noise, but at the same time will shorten unambiguous range of measurements.

3.2 Kinect V2 Features

The Kinect V2 diverges from the textbook case by acquiring three samples of the correlation function with suppressed background illumination with phase offsets of $0, \frac{2\pi}{3}, \frac{4\pi}{3}$ as described in patent[11], with a formula different to equation 3.

$$\phi = atan2(-\sum_{k=1}^{3} \tau_k * sin(\frac{2\pi}{3} \times (k-1)), \sum_{k=1}^{3} \tau_k * cos(\frac{2\pi}{3} \times (k-1))) \tag{5}$$

According to the patent this reduces the effect of varying temperature, the influence of a imperfect modulation signal and the variation of the components of the device over time. The required amount of calibration data per pixel is reduced to the phase offset induced by the signal propagation delay of the hardware.

Three modulation frequencies are used in order to get unambiguous measurements in the standard living room scenario while still achieving a high depth resolution. The frequencies being used are approximately 16 MHz, 80 MHz and 120 MHz, yielding an unambiguous range of 9.37 meters for the lowest frequency or 18.74 meters with frequency based phase unwrapping as described by Dröschel et al.[12]. Multiple modulation frequencies allow identification of errors due to multiple returns at single pixels, induced by scattering of light within the scene or at edges of objects.

Additionally, as with the Kinect V1, the depth sensor is accompanied by a regular RGB camera. Its resolution has been increased to 1920x1080 pixels, compared to 640x480 pixels of the predecessor.

3.3 Other depth cameras

Name	Range (m)	Resolution	Framerate	Precision(mm / 1σ)	FOV (degrees)
Kinect V2	not specified	512 × 424	30	not specified	89 × 71
pmd CamBoard nano	0.5	160 × 120	90	not specified	90 × 48
Kinect V1	4	640 × 480	30	50 @ 4m	57 × 43
SoftKinetic DS325	not specified	320 × 240	30	14 @ 1m	74 × 58
pmd CamCube	7	200 × 200	40	3 @ 4m	40 × 40
SwissRanger SR4000	7	176 × 144	50	4 @ 2m	43 × 34

There is a broad range of depth cameras available today for different scenarios. All cameras are based on continuous modulation, except for the Kinect V1, which projects structured light, performs stereo matching and thus does not offer per pixel depth measurements. The first 3 devices are consumer oriented, having prices between 200 and 690$. The last 2 devices are designed for industrial applications with a reduced field of view, higher precision and range and prices tags at around 5000$. The modulation frequency typically is in the range of 10 to 80 MHz. There are a few depth cameras that have identical sensors as the Kinect V1 in terms of depth imaging. For our experiments we used a Carmine 1.08, does not require on an additional power supply.

4. EVALUATION AND RESULTS

4.1 Accuracy on calibration target

We detected a $1m^2$ square print of fiducial markers[13] in the intensity image to estimate the position and orientation of the marker board. The intrinsic camera parameters could be verified based on the average reprojection error being below 0.4 for all detections. Partial groundtruth depth images were generated based on the known pose and geometry of the board. Data was recorded in a tight office space and in a wide attic. The first location was chosen to possibly observe the multipath effect in a standard use case with light walls or other reflective surfaces, the second location was chosen to avoid these error sources.

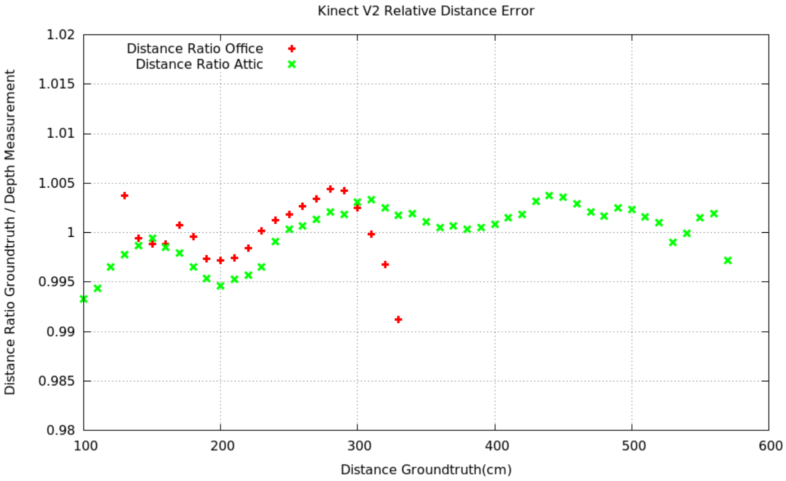

Figure 2. Average pixelwise ratio of groundtruth and measured depth of a calibration target at different distances. The data was recorded in a tight office space as well as in a very wide attic. A significant depth based bias can be observed. At 5.5 meters the marker board was too small to be reliably detected.

The ratio of groundtruth depth to measured depth at different distances was plotted in Figure 2. Only pixels showing white parts of the marker board are evaluated. The dominating error source in the attic plot is a systematic distance dependant error. One possible error source is the non-ideal optical modulation signal according to Schmidt and Jähne.[14] The office plot shows a similar pattern with additional errors, due to multipath effects. In both cases the error stays below 1% over a range of 5.5 meters. At this distance the board was to small to be detected in the IR image.

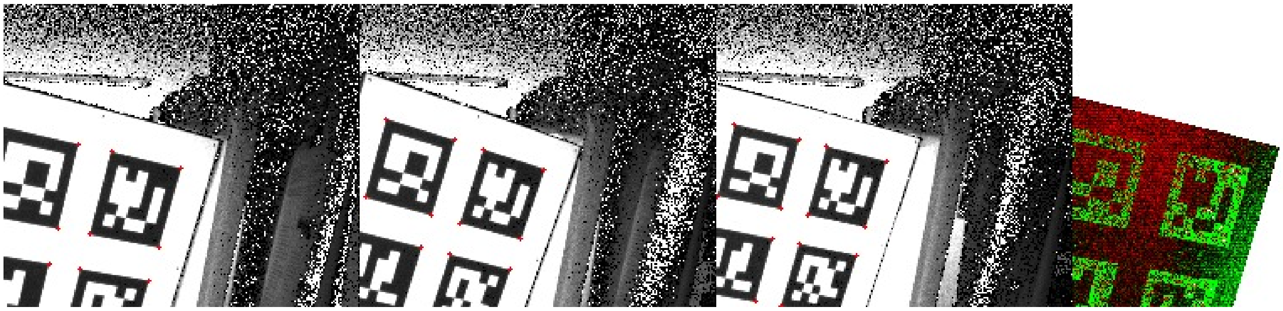

Figure 3. Reflections of the whiteboard influence measurements on the marker board and vice versa. The noise on the whiteboard moves according to the position of the marker board as can be seen in the first three images. The relative distance error has been visualized in the last image. Overestimated distances are shown in green, underestimated distances in red. Dark parts of the board show a intensity related depth bias.

One hour of data of a completely static setup of sensor and marker board was recorded in order to observe measurement stability over time. Figure 4 shows that with increasing temperature, the measured depth increases as well. After approximately 1000 seconds the device activates an internal cooling fan. Its intended use seems to be to prevent the device from overheating, rather than ensuring stable measurements. An additional temperature sensor or an object of known geometry in the scene are required, in order to compensate for this effect.

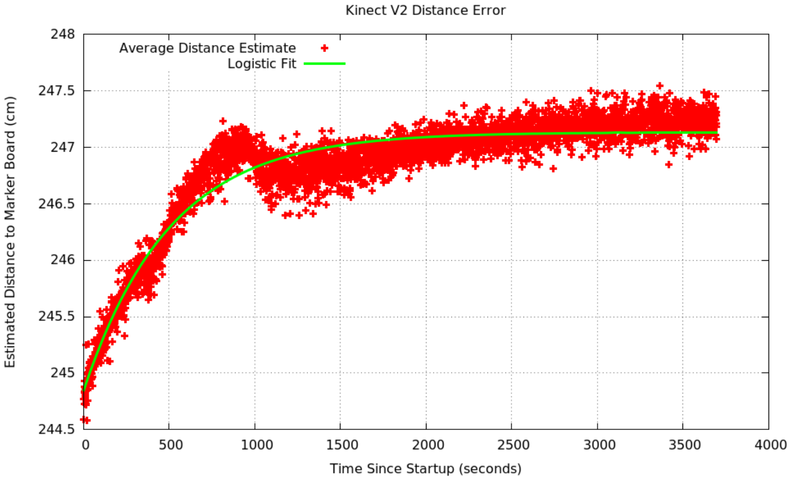

Figure 4. One hour of depth measurements of a stationary marker board show a temperature drift. After around 1000 seconds a cooling fan is activated, delaying the point of thermal equilibrium.

4.2 Precision on a planar surface

Depth data of a planar wall was recorded with the Kinect V2 and a PrimeSense Carmine 1.08. 1000 frames per device were saved and their average was calculated in order to get depth images with low noise. In the case of the Carmine this was done by averaging over the depth values, in the Kinect V2 case raw sensor data was averaged. We estimated the orientation of a plane using RANSAC with an inlier threshold of 2 cm and got inlier sets sufficiently large to perform a plane fitting test comparable to Khoshelham[3]. Since both sensors had a slanted view on the wall, a depth error in respect to the distance from device could be determined from a single view.

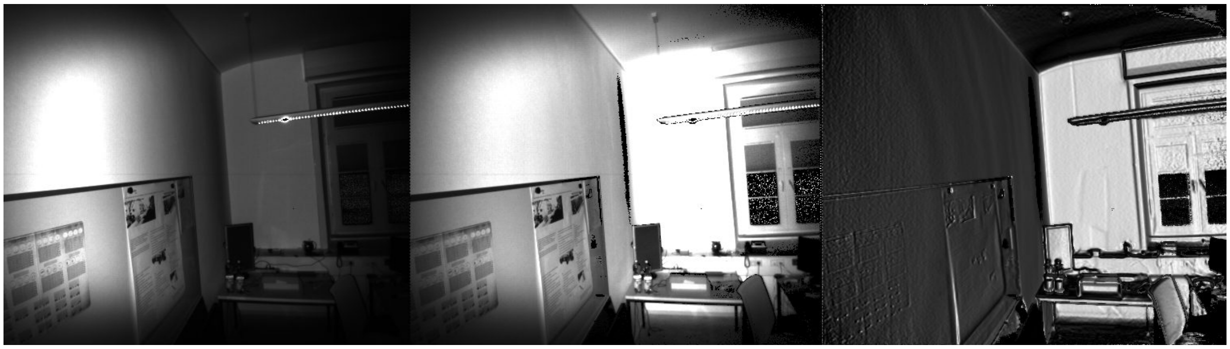

Figure 5. Planar wall recorded for the plane fitting test. The images are an IR intensity image truncated to 8 bit showing the illumination falloff with increasing distance, illumination normalized by distance and a visualization of the estimate surface normal of every pixel. The intensity is defined as the euclidean norm of two corresponding correlation coefficients.

We performed three different approaches to calculate the depth error:

- Standard deviation: The quadratic mean of differences of depth measurements and their average of 1000 frames.
- Euclidean distance to plane: The quadratic mean of differences of depth measurements and their closest points on the estimated plane. Same metric as in Khoshelham[3].
- Depth distance to plane: The quadratic mean of differences of depth measurements and their expected depth based on the estimated plane.

With n being the normal vector of the plane, r being its distance to the origin and p being a vector describing the coordinate of a point corresponding to a depth measurement, the euclidean distance d from point to plane is calculated as given in Equation 6.

$$d = \|n \cdot p - r\|_2 \tag{6}$$

The depth distance z to the plane is the difference of the z components of the measured point p and the intersection point p' of the plane and a line going through the origin and p:

$$\begin{gathered} p' = x \times p, 0 = n \cdot p' - r \\ z = \left\| (p - p') \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\|_2 \end{gathered} \tag{7}$$

The errors of the PrimeSense Carmine depth measurements are plotted in Figure 6. The fit to the standard deviation is nearly identical to the result in Khoshelham[3], showing the precision in respect to the measuring distance. Although the euclidean distance to the estimated plane is the distance used in Khoshelham[3], it underestimates the true error for surfaces not perpendicular to the viewing direction. The z distance corresponds to the difference of an ideal depth measurement of a single point and the actual depth measurement performed by the device. As the other errors it is influenced by noise and additionally suffers from a bias for depth estimation on slanted surfaces, introduced by the stereo matching with a large window size.

Figure 7 shows the pixelwise errors of the Kinect V2 depth measurements. The standard deviation is slightly higher, even under ideal conditions as high reflectivity and no multipath errors. The difference between standard deviation and z distance is smaller compared to results shown in Figure 6, since the distances is measured by every pixel independently.

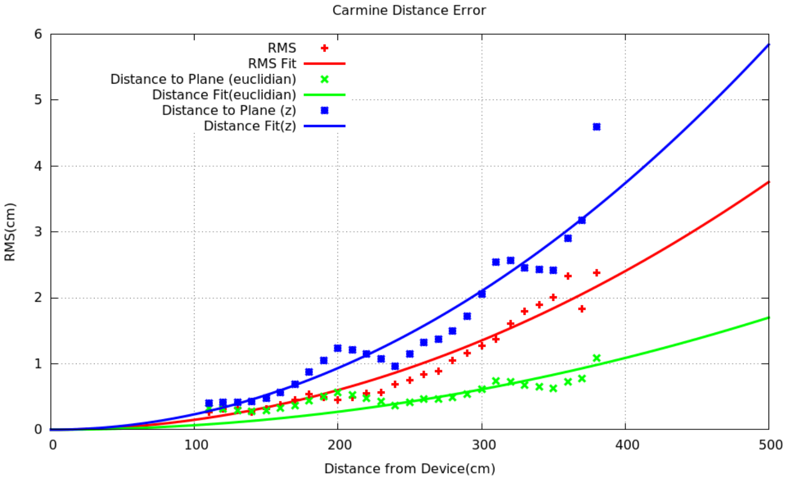

Figure 6. Average error in respect to the measuring distance for a PrimeSense Carmine. The error increases proportional to the square of the distance. The RMS fit is almost identical to the one in Khoshelham[3] with the coefficient being 1.503×10^{-5}. The higher number of samples shows some anomalies in the z distances, probably due to depth quantization or sparseness of the projected point pattern. The ratio between standard deviation and RMS depth error is 0.64

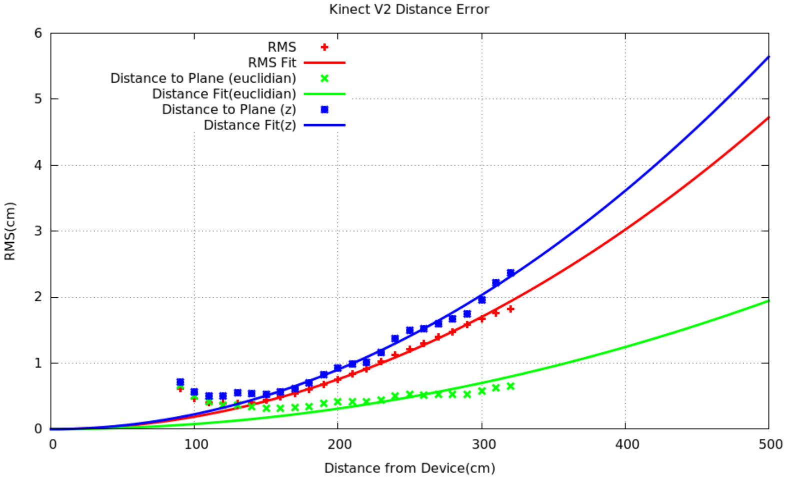

Figure 7. Average error in respect to the measuring distance for the Kinect V2. The error increases more regularly compared to the results in Figure 6. At short distances the error increases slightly, due to a illumination falloff at the border of the image. The ratio between standard deviation and RMS depth error is 0.84.

In order to have a fair comparison of both sensors we performed bilateral filtering on the correlation coefficients, the sums of Equation 5, on a 3×3 neighbourhood with the results shown in Figure 8. As can be seen, the standard deviation is reduced to 36%, while the average z distance is only reduced to 57%. Smoothing the correlation coefficients leads to a higher signal to noise ratio while leading to a lower angular resolution and a lower ratio between standard deviation and average depth error.

4.3 Angular Resolution

The Kinect V2 has a horizontal field of view of approximately 90°. At 512 pixels resolution this yields an angular resolution of 0.176°, while the Kinect V1 has 57° field of view, 640 pixels resolution and thus 0.089° of angular resolution. Since Martinez[4] suggests that stereo matching in the Kinect V1 is performed over a windows of 17×17 pixels in the IR image, every depth measurement is likely to be correlated to its 7×7 neighbourhood,

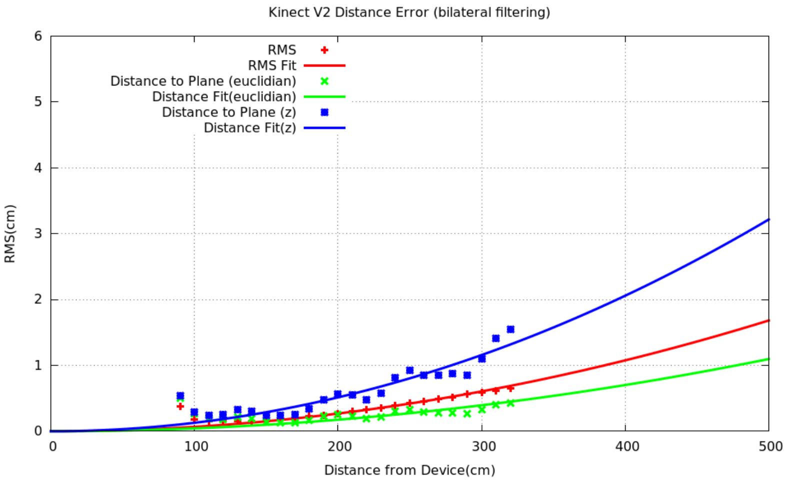

Figure 8. Errors as in Figure 7 with bilateral filtering enabled. The errors drop significantly by averaging over a 3 × 3 neighbourhood and thus increasing the signal to noise ratio. The ratio between standard deviation and RMS depth error is 0.52

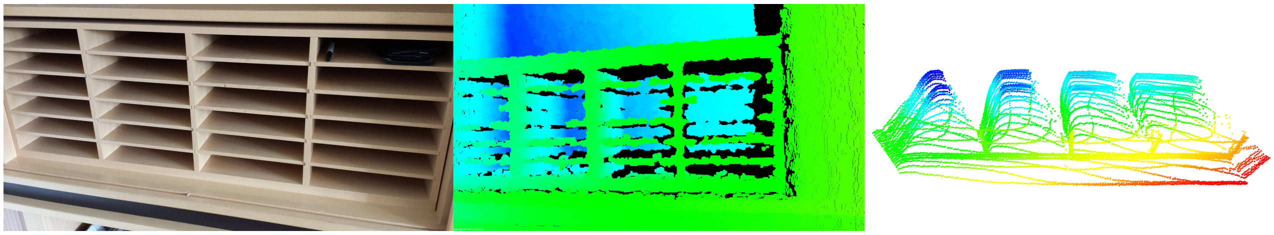

Figure 9. A cubby hole shelf was recorded with a Carmine PrimeSense (middle image) and a Kinect V2 (right image). The Carmine states the depth of the thin boards at either the foreground or the background depth in an irregular way. The Kinect V2 measurements show strong multipath in the back of the compartments and flying pixels at the fine structures.

thus reducing the angular resolution. In order to illustrate this aspect, we recorded depth images of a cubby hole shelf with both sensors. The thin wooden boards separating two vertically adjacent compartments are 8mm thick and have a height of 2 - 3 pixels in the Kinect V2 and 5 pixels in the Carmine depth image.

Figure 9 shows a visualization of point clouds acquired with the Kinect V2 and a Carmine. The Carmine struggles at resolving the fine structure of the shelf showing a mix of correct and wrong depth estimates as well as some holes with no valid depth estimate. The sparseness of the random dot pattern used for stereo matching yields an irregular pattern in the depth image. The Kinect V2 depth image does not show any holes or irregular depth patterns. Instead it shows ”flying pixels” at pixels measuring two returns at once. They are forming S-shaped traces of points as single pixel rows fade from viewing foreground to background. The flying pixel effect is more pronounced when the foreground and background reflect a similar amount of light (e.g. same material and distance). In a gesture recognition system with skin in the foreground (high infrared albedo) and background at a greater distance, this effect will be less problematic (see Figure 10).

5. OUTLOOK

We provided a detailed analysis of the Kinect V2’s accuracy and precision in standard scenarios. It was shown that the new sensor shows similar characteristics as its predecessor in terms of precision that is proportional to the squared measuring distance. Experiments with a calibration target show a significant distance dependant error, requiring further calibration. Although the image resolution of the depth camera has been decreased, the device is capable of better resolving fine structures in certain scenarios. Some aspects of the devices require further investigation. The timing of the acquisition of correlation images is unknown. With this information

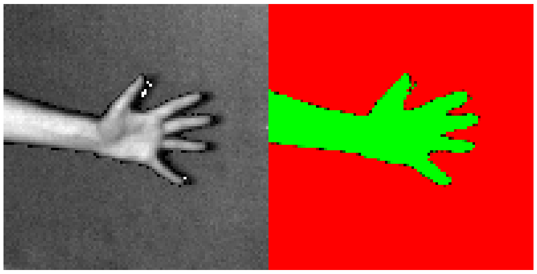

Figure 10. Intensity and depth image of a hand against a dark background. Due to the strong contrast a detailed silhouette can be seen.

a correction of motion artifacts could be possible in certain situations. The warm up experiment suggests a temperature dependant bias that needs to be corrected for. We do not now if any helpful information is being delivered by the device, as some parts of the communication protocol may still be unknown. The open source driver allows many interesting applications that encourage for more experimentation. Integration over multiple raw frames has already been implemented, offering measuring ranges of up to 18 meters. Analysis of pixel noise shows a strong relation to non-correlated scene illumination (e.g. sunlight), suggesting the deduction of a passive infrared frame for registration with other cameras. The high resolution RGB camera provides a second image for triangulation, offering an extended range and higher accuracy in short range measuring.

ACKNOWLEDGMENTS

We would like to thank the team around libfreenect2, that implemented free Kinect V2 driver giving access to functionality not included in the official SDK.

REFERENCES

[1] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R., "Real-time human pose recognition in parts from single depth images," *Communications of the ACM* **56**(1), 116–124 (2013).

[2] Newcombe, R. A., Davison, A. J., Izadi, S., Kohli, P., Hilliges, O., Shotton, J., Molyneaux, D., Hodges, S., Kim, D., and Fitzgibbon, A., "Kinectfusion: Real-time dense surface mapping and tracking," in [*Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*], 127–136, IEEE (2011).

[3] Khoshelham, K., "Accuracy analysis of kinect depth data," in [*ISPRS workshop laser scanning*], **38**(5), W12 (2011).

[4] Martinez, M. and Stiefelhagen, R., "Kinect unleashed: Getting control over high resolution depth maps,"

[5] Hirschmuller, H., "Accurate and efficient stereo processing by semi-global matching and mutual information," in [*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*], **2**, 807–814, IEEE (2005).

[6] Lefloch, D., Nair, R., Lenzen, F., Schäfer, H., Streeter, L., Cree, M. J., Koch, R., and Kolb, A., "Technical foundation and calibration methods for time-of-flight cameras," in [*Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*], 3–24, Springer (2013).

[7] Lindner, M. and Kolb, A., "Lateral and depth calibration of pmd-distance sensors," in [*Advances in Visual Computing*], 524–533, Springer (2006).

[8] Kahlmann, T., Remondino, F., and Ingensand, H., "Calibration for increased accuracy of the range imaging camera swissrangertm," *Image Engineering and Vision Metrology (IEVM)* **36**(3), 136–141 (2006).

[9] Schiller, I., Beder, C., and Koch, R., "Calibration of a pmd-camera using a planar calibration pattern together with a multi-camera setup," *The international archives of the photogrammetry, remote sensing and spatial information sciences* **37**, 297–302 (2008).

[10] Lindner, M., Schiller, I., Kolb, A., and Koch, R., "Time-of-flight sensor calibration for accurate range sensing," *Computer Vision and Image Understanding* **114**(12), 1318–1328 (2010).

[11] Xu, Z., Perry, T., and Hills, G., "Method and system for multi-phase dynamic calibration of three-dimensional (3d) sensors in a time-of-flight system," (Nov. 19 2013). US Patent 8,587,771.

[12] Droeschel, D., Holz, D., and Behnke, S., "Multi-frequency phase unwrapping for time-of-flight cameras," in [*Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*], 1463–1469, IEEE (2010).

[13] Munoz-Salinas, R., "Aruco: A minimal library for augmented reality applications based on opencv," (2012).

[14] Schmidt, M. and Jähne, B., "A physical model of time-of-flight 3d imaging systems, including suppression of ambient light," in [*Dynamic 3D Imaging*], 1–15, Springer (2009).