On Maximum Geometric Finger-Tip Recognition Distance Using Depth Sensors

Marius Shekow, Leif Oppermann
Fraunhofer FIT
Schloss Birlinghoven
53754 Sankt Augustin, Germany
(marius.shekow|leif.oppermann)
@fit.fraunhofer.de

ABSTRACT

Depth sensor data is commonly used as the basis for Natural User Interfaces (NUI). The recent availability of different camera systems at affordable prices has caused a significant uptake in the research community, e.g. for building hand-pose or gesture-based controls in various scenarios and with different algorithms. The limited resolution and noise of the utilized cameras naturally puts a constraint on the distance between camera and user at which a meaningful interaction can still be designed for. We therefore conducted extensive accuracy experiments to explore the maximum distance that allows for recognizing finger-tips of an average-sized hand using three popular depth cameras (SwissRanger SR4000, Microsoft Kinect for Windows and the Alpha Development Kit of the Kinect for Windows 2), with two geometric algorithms and a manual image analysis.

In our experiment, the palm faces the sensors with all five fingers extended. It is moved at distances of 0.5 to 3.5 meters from the sensor. Quantitative data is collected regarding the number of finger-tips recognized in the binary hand outline image for each sensor, using two algorithms. For qualitative analysis, samples of the hand outline are also collected.

The quantitative results proved to be inconclusive due to false positives or negatives caused by noise. In turn our qualitative analysis, achieved by inspecting the hand outline images manually, provides conclusive understanding of the depth data quality. We find that recognition works reliably up to 1.5 m (SR4000, Kinect) and 2.4 m (Kinect 2). These insights are generally applicable for designing NUIs that rely on depth sensor data.

Keywords

Natural User Interaction, Depth sensor, Finger-tip recognition, SwissRanger SR4000, Microsoft Kinect, Kinect for Windows 2 alpha development kit

1 INTRODUCTION

In Human Computer Interaction, mouse, keyboard and touch screens are today's standard input methods. As the user's interaction space is limited due to being bound by physical contact, Natural User Interfaces (NUI) have gained popularity in the research community. Hand gesture interfaces are an important NUI branch and recently consumer-grade applications have emerged[1] which are based on depth cameras. Such vision-based approaches allow for non-intrusive interaction, where no physical contact between user and device is required. Generally solutions are often based on inexpensive commodity color cameras (e.g. based on CMOS or CCD technology) with high resolution. Their down-side is that segmenting the image and analyzing its content is computationally challenging and negatively affected by varying lighting conditions. Thus, many solutions instead use depth sensors which allow for simple, thresholding-based scene segmentation. Unfortunately the depth data quality is negatively affected by noise and low image resolution. These issues are a challenge for recognition algorithms, especially at higher distances where objects decrease in size due to perspective foreshortening. Therefore we investigate the sensor-specific maximum hand distance where hands and finger-tips are still recognizable. We are not concerned about the *minimum* finger-tip recognition distance as it relates to the minimum distance for which a camera can report depth values[2].

[1] Examples include Intel Perceptual Computing SDK `http://software.intel.com/en-us/vcsource/tools/perceptual-computing-sdk`, Leap Motion `https://www.leapmotion.com` or various hand and body gesture recognition tools based on SDKs like the Microsoft Kinect SDK `http://www.kinectforwindows.org` or the SoftKinetic iisu `http://www.softkinetic.com/en-us/products/iisumiddleware.aspx`.

[2] At low distances, perspective foreshortening is no longer an issue and noise can be effectively dealt with using various filtering techniques.

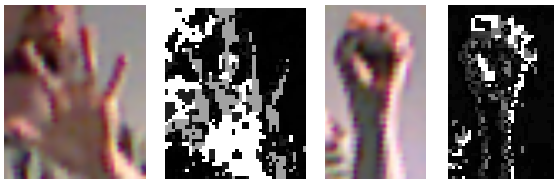


Figure 1: Hand segmentation by skin-color classification using LCCS method

The rest of this work is structured as follows. Section 2 presents related work, together with background information regarding depth cameras and hand gesture recognition algorithms. The geometric algorithms that detect finger-tips are explained in section 3. The experiment and its results are found in section 4 and 5 respectively. A conclusion of this work is given in section 6.

2 BACKGROUND & RELATED WORK

Two prominent depth sensing technologies used by sensors found on the market are Time-Of-Flight (TOF) and Infrared-Structured-Light (IRSL). TOF cameras have only become affordable recently[3] while IRSL was established on the market by PrimeSense, with vendors such as the Microsoft (Kinect) and ASUS (Xtion). The disadvantage of depth sensors is their relatively low resolution, often around 250x250 px for TOF or 320x240 px for IRSL. Sensors like the Microsoft Kinect provide depth and RGB data, but it should be noted that obtaining the hand outline by performing skin-color classification on the RGB data is very challenging in real-world environments. We found that pixels of (parts) of the hands are represented by RGB colors that are not of skin color, causing any parametric or non-parametric method to fail. Results of performing skin color classification using log-chromatic color space (LCCS) [8] on RGB images obtained using the Kinect can be seen in figure 1.

Various classes of hand gesture recognition algorithms exist, see [4], such as appearance-based, partial and full-DOF pose estimation. For any of these classes the quality of incoming data, depth values in our case, is of crucial importance. Hand pose estimation analyzes the pixels belonging to hands and fingers, which is challenging due to the low depth image resolution (see table 1) and due to perspective projection causing the size of objects to shrink in the image with increasing distance. If data is missing, e.g. an individual finger that disappears in the depth image, any recognition algorithms will intermittently fail. For this reason this paper provides an analysis regarding the discernibility of individual fingers in the depth image produced by the SwissRanger SR4000 (SR4k) and Microsoft Kinect for Windows (K4W), which are sensors often used by hand gesture recognition researchers. For an outlook, results are also provided for the alpha development kit of the Microsoft Kinect for Windows 2 (K4W2alpha), the retail version[4] of which being expected in the second half of 2014.

Related work is sparse which measures the feasibility to detect the number of fingers in a depth image in relation to the hand distance. The majority of works (e.g. [1, 7, 9, 10, 12, 13]) examine the noise characteristics, that is, precision, jitter, repeatability, etc. of reported Z values of a depth sensor, some also investigate lens distortion. Mishra et al. [11] do examine the X-Y accuracy by using a monkey wrench. For different aperture sizes ranging from 0.2 to 1.6 cm, the authors measured the maximum distance for which the aperture hole would still be visible. Unfortunately, the gap between finger-tips exceeds the range used in their experiments (finger-to-finger distance being ~ 2.5 cm).

We therefore designed an experiment where we present a single hand with 5 fingers, the palm facing the camera, at varying distances between 0.5 and 3.5 meters. We use the 3DMT [5] implementation which uses this 5-finger pose to locate the hand, providing tracking for the hand and fingers. By extending 3DMT with a second algorithm, convexity defects, the number of fingers are determined by both algorithms in parallel, logging the results to disk. The two algorithms analyze the hand geometrically and are briefly introduced in section 3. In addition to this quantitative analysis, the discernibility of fingers is also evaluated qualitatively by judging whether the binary hand silhouette exhibits all 5 fingers clearly. The result serves as guideline for hand gesture researchers who want to be informed about the limitations of common depth sensor hardware.

3 ALGORITHMS

In this work two partial pose estimation algorithms were used to determine the number of fingers. In partial pose estimation, a simple hand model is used which consists of the palm (approximated by a circle) and finger-tips, as well as finger-pipes in some cases. The two algorithms used here are *3D multi touch* (3DMT) and *convexity defects* (CD), presented in the following subsections. Both analyze the pixels with a *geometric* approach to extract the palm and finger-tips.

3.1 3D Multi Touch

The 3DMT toolkit [5, 6] was developed by Georg Hackenberg in 2010 at Fraunhofer FIT. It extracts from the depth image a hand model that consists

[3] In 2008 the SR4k costed $\sim$ 10000€ whereas today similar products, such as the SoftKinetic DS311 (< 300 USD) or the Creative Senz3D ($\sim$ 175 USD) are much less expensive.

[4] Improvements of the depth data quality between the alpha development kit and the retail version are possible.

	SR4k	K4W	K4W2alpha
Techn.	TOF	IRSL	TOF
Price	∼7000 €	∼190 €	N/A
Range	0.1 - 5.0 m	0.4 - 4.0 m	0.5 - 4.5
Depth res.	172x144	640x480	512x424
FOV (H, V)	43.6°, 34.6°	58°, 45°	84.1°, 53.8°
Refresh rate	∼30 Hz	30 Hz	30 Hz
Outdoor usage	yes	no	yes

Table 1: Depth sensor technical specifications

Adapted from [2], the SR4k data sheet and the preliminary K4W2alpha technical specification sheet.

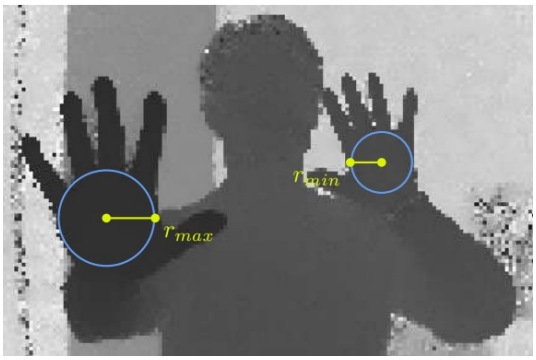
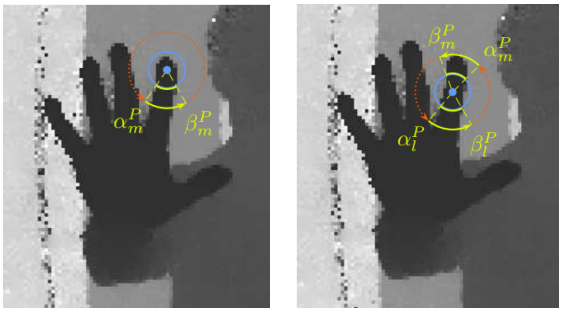


Figure 2: Palm radius determination by [6]



(a) Finger-tip circle check (b) Finger-pipe circle check

Figure 3: Finger-tip and -pipe checking

of a palm and 0-5 fingers. In this model a finger is made of a finger-tip which is connected to the palm via a finger-pipe. To determine the dimension of the palm, a circle is placed around the preliminary center (based on distance-transform) of each hand. Its radius is iteratively increased, until the depth values along the circle are no longer approximately equal to the center's depth value, see figure 2. The next step is to find finger-tips by finding pixels that are finger-tip candidates. A fixed-circle test is passed if the depth values along the circle can be grouped in two segments, one where the values deviate strongly from the one of the center and one where the deviation is low (below a threshold), see figure 3a. A similar fixed-size circle test is performed for computing candidates for finger-pipes, see figure 3b. This results in a map indicating the presence of finger-tips and pipes respectively. False positives are eliminated through smoothing, and finger-tip candidates are connected to palms via finger-pipes wherever it is geometrically plausible. The result can be seen in figure 4.

The original work was modified to support the K4W and K4W2alpha sensors in addition to the SR4k. The algorithm was also made scale-adaptive, s.t. the experimentally determined parameters (regarding the circle checks, distance thresholds, etc.) are adjusted according to the hand distance.

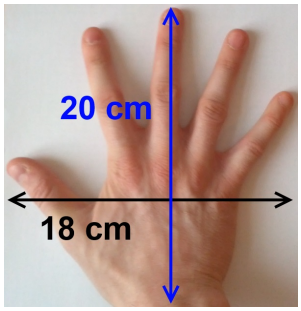
Figure 4: Finger-tips and -pipes

3.2 Convexity Defects

"Convexity defects (CD) in a (convex) hull is the space between the contour line and the actual object." [14] This is best illustrated in figure 5. It shows an *object* (white), its *contour* or *outline* (red), *convex hull* (green) and the *convexity defects* (brown). An efficient implementation is provided by OpenCV [3] for determining convexity defects. While the 3DMT finger recognition algorithm operates directly on the depth image, CD needs the hand outline. Therefore, depth thresholding is applied to obtain the binary hand silhouette from

Figure 5: Convexity defects

(a) Sensor setup



(b) Hand pose and dimensions

Figure 6: Experiment setup

which the contour, convex hull and CD's are extracted. OpenCV describes the CD area in terms of the outline, providing the start and end index, which coincide with the finger-tips. To remove false positives around the wrist, only those points which are above the palm center (in image space) are considered to be finger-tips.

The 3DMT implementation was extended to carry out the CD-based finger-tip extraction in parallel to the original 3DMT algorithm. Finally, logging functionality was added to write tuples (hand distance, 3DMT-fingers, CD-fingers) to disk.

4 EXPERIMENT

All 3 sensors were positioned and oriented so that they capture approximately the same environment, see figure 6a. One hand was presented to the sensors with all 5 fingers extended, as shown in figure 6b. The 3DMT implementation comes with recording functionality where the incoming depth frames are stored to disk *without* applying the recognition algorithm. This rules out runtime performance differences that arise when executing recognition on depth frames of varying resolution (see table 1). The K4W was operated at 320x240 pixels instead of 640x480, because we found that the higher resolution does not improve the level of detail, as can be seen in figure 7. While recording, the hand is initially placed at a distance of 1.0 m and then moved towards the camera until the respective minimum distance of the sensor, followed by moving it away to a distance of up to 3.5 m, see figure 8. The SR4k and K4W2alpha sensors cannot be operated simultaneously due to infrared inference, thus only one device was active at a time while recording. The hand movement sequence was executed twice to obtain more data, resulting in two recordings for each sensor. While replaying the recordings, the 3DMT and CD algorithms were applied while logging the results to disk.

(a) 320x240 (b) 640x480

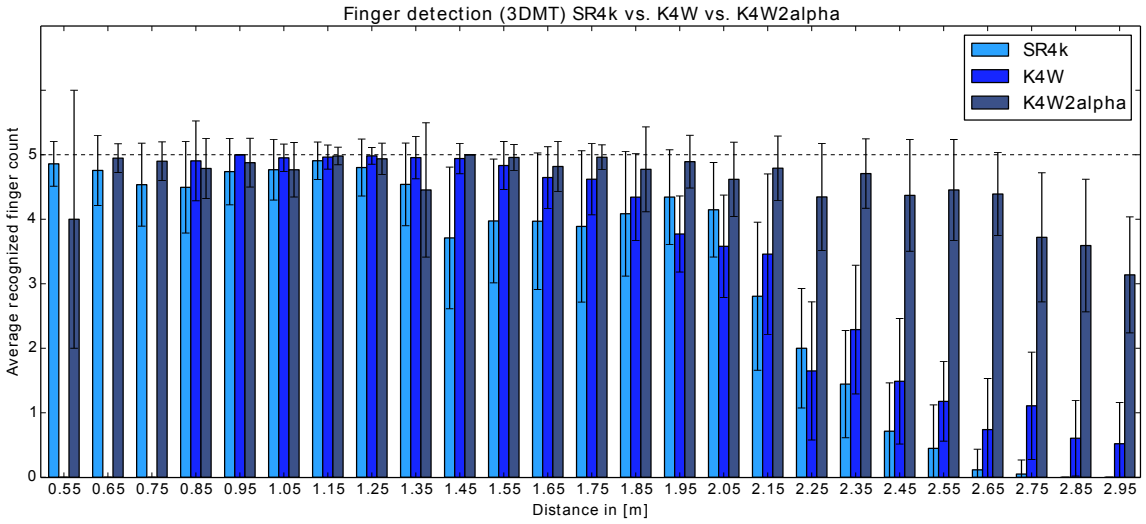
Figure 7: K4W depth data resolution comparison (hand at a distance of 2.4 m)

5 RESULTS

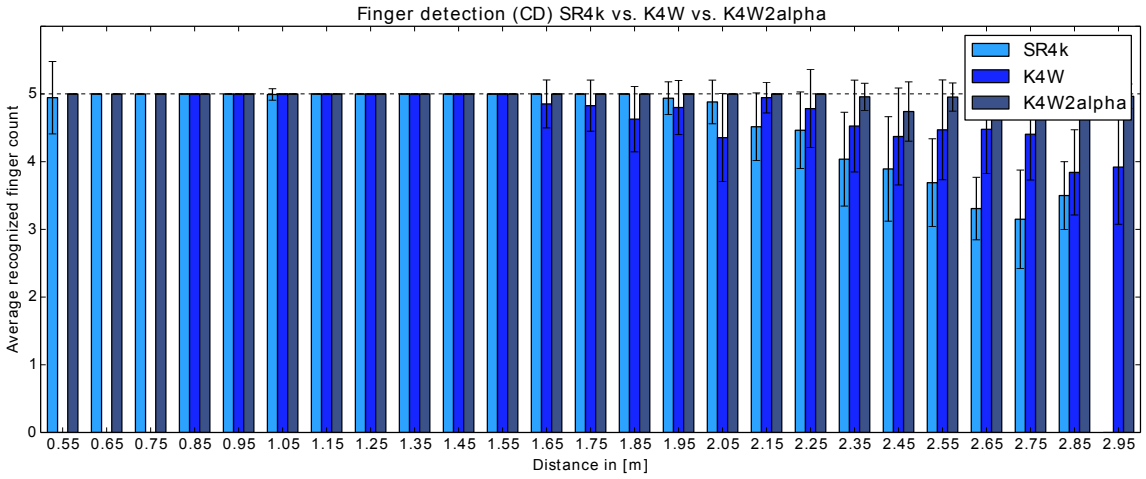
We first present our quantitative results, represented by the bar plots shown in figure 9. It contains plots of the average number of detected fingers together with the standard deviation, in relation to the hand distance, binned in 10 cm intervals. Only the K4W2alpha and SR4k provide values for distances below 0.8 m. Although the K4W sensor does support a minimum distance of 0.4 m in *near-mode*, its use was discarded because it has a negative effect on depth image quality for objects at larger distances (beyond 1 m), compared to normal mode. The performance of the 3DMT algorithm for the SR4k at near (1.0 m) to mid distances (1.5 m) is below par (i.e., less than 5 fingers are recognized) due to noise leading to incoherent finger-pipe maps. If no sufficiently large, continuous finger-pipe blob between a finger-tip *candidate* and the palm is found, the finger-tip is not counted as detected finger.

The bar plots alone do not provide a conclusive answer to the aims of this work. In theory, one could conclude that the sensor's depth data quality is sufficient to allow the recognition of all fingers for a certain distance if the respective bar shows a perfect finger count in both figure 9a and 9b. In practice, we found that the 3DMT

Figure 8: Snapshots of the evaluation sequence (SR4k)



(a) Results (3DMT algorithm)



(b) Results (convexity defects algorithm)

Figure 9: Quantitative analysis results

algorithm undershoots, while the CD algorithm overshoots, due to random noise that is present for every sensor. Exemplary, according to CD, the K4W2alpha is able to find nearly all fingers at a distances of 3.0 m, which is incorrect, as we will show below. Discarding CD as inappropriate measure, we can still use 3DMT and define an (arbitrary) lower-bound threshold of, say, 4.5 fingers, in which case the maximum distances were as follows:

- SR4k: 1.4 m
- K4W: 1.8 m
- K4W2alpha: 2.6 m

In order to make a conclusive statement a qualitative, manual analysis was performed on the image data. A sample is presented in table 2 for each sensor. It shows binary hand silhouettes extracted from the depth image using a simple thresholding of ± 9 mm around the palm center. It should be noted that 3DMT and CD apply light filtering to the input to reduce the effects of noise, such as Median smoothing and morphological closing. The images in table 2 do not contain this filtering to allow for a more distinct judgment. They are scaled to a uniform size and include the width of the hand in pixels in the original, unscaled image, which relates to the width of the real hand of 18 cm.

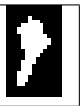
Distance [m]	0.6	0.8	1.2	1.5	1.8	2.1	2.3	2.7	3.0
SR4k									
Width [px]	69	55	36	30	25	20	16	10	9
K4W	N/A								
Width [px]	N/A	60	40	34	26	24	21	16	16
K4W2alpha									
Width [px]	110	85	58	46	40	33	30	24	21

Table 2: Qualitative analysis results: binary hand outlines at different distances

From the image analysis we find the following maximum distances where all five finger-tips are visible:

- SR4k: 1.5 m
- K4W: 1.5 m
- K4W2alpha: 2.4 m

Note that these numbers result from analyzing *several* hand images with approximately the same distance to determine jitter in the images, such as disappearing fingers. Exemplary, the hand at distance 1.8 m in the SR4k row in table 2 looks correct and thus one might assume that the SR4k supports finger-tip recognition at this distance. Unfortunately, at such a high distance fingers become 1 px thin and begin to disappear sporadically. At a distance of 1.5 m finger-tips have a stable appearance. Table 3 shows a sample of hand outline images, all taken at the respective maximum distance for each sensor, corresponding to the list above. Its last row exemplifies that the result of 2.6 m for the K4W2alpha from the quantitative analysis is not appropriate, as fingers sporadically disappear.

6 CONCLUSION

In this work we tested the appearance of hands and finger-tips in depth images produced by three depth sensors frequently used in hand gesture research, the MesaImaging SwissRanger SR4000 (released in 2008), Microsoft Kinect for Windows (released in 2010 [5]) and the alpha development kit of the Kinect for Windows 2 (retail version expected Summer 2014). In an experiment we presented a single hand with all 5 fingers extended to the sensors, at hand-to-sensor distances ranging from 0.5 to 3.5 meters. We used the 3DMT toolkit and its recognition algorithm to find the hand and its finger-tips. We extended it with a second geometrical algorithm, convexity defects, and added disk-logging of hand distance and number of found finger-tips. This quantitative data was summarized in histograms, see figure 9.

Comparing the algorithms we found that the convexity defects algorithm consistently overshoots regarding the number of found finger-tips, due to noise at increasing distances. This is also confirmed by the authors in [4], suggesting the use of local palm-to-finger-tip distance maxima in future work. The 3DMT algorithm instead provides a conservative estimation, i.e., undershoots the true number of recognizable finger-tips.

To make a conclusive statement an additional qualitative analysis was performed on the image data, see table 2 and 3. We find that recognition is possible up to 1.5 m with the SwissRanger SR4000 and Kinect for Windows and up to 2.4 m for the alpha development kit of the Kinect for Windows 2. The main advantages of the new Kinect over its predecessor are the depth measurement technology (Time-Of-Flight vs. Infrared-Structured-Light) and high resolution (in the realm of Time-Of-Flight cameras) of 512x424 pixels. With depth sensor resolutions improving in the future, one can expect a larger interaction space for hand gesture applications. A confirmation of the results with the retail version of this sensor is left as future work.

The main contribution of this work is a method to determine the maximum hand distance that allows for geometric finger-tip recognition using depth cameras, including its limitations when using only quantitative data. We hope that authors pick up on our work using extended parameters, such as testing different sensors, varying sunlight conditions or differently sized hands. Our results provide hand gesture researchers with an orientation and guideline regarding the finger-tip recognition capabilities of present depth sensor hardware.

[5] The XBox 360 Kinect was released in 2010, whereas the Kinect for Windows was officially released in 2012. The underlying technology is the same, IRSL by PrimeSense.

SR4k (1.5 m)	
K4W (1.5 m)	
K4W2alpha (2.4 m)	
K4W2alpha (2.6 m)	

Table 3: Qualitative analysis results: binary hand outlines at maximum distance

Note that for K4W2alpha, *this is preliminary software and/or hardware and APIs are preliminary and subject to change*.

7 REFERENCES

[1] M. R. Andersen, T. Jensen, P. Lisouski, A. K. Mortensen, M. K. Hansen, T. Gregersen, and P. Ahrendt. *Kinect Depth Sensor Evaluation for Computer Vision Applications: Technical Report ECE-TR-6*. 2012.

[2] Arne Bernin. Einsatz von 3D Kameras zur Interpretation von räumlichen Gesten im Smart Home Kontext. Master's thesis, Hochschule für angewandte Wissenschaften Hamburg, Germany, 2011.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1-2):52–73, 2007.

[5] G. Hackenberg, R. McCall, and W. Broll. Lightweight palm and finger tracking for real-time 3D gesture control. In *Virtual Reality Conference (VR), 2011 IEEE*, pages 19–26, 2011.

[6] Georg Hackenberg. Development of a Multi-Touch Interface using a 3D Camera. Master's thesis, RWTH Aachen, Germany, 2010.

[7] H. Haggag, M. Hossny, D. Filippidis, D. Creighton, S. Nahavandi, and V. Puri, editors. *Measuring depth accuracy in RGBD cameras: Signal Processing and Communication Systems (ICSPCS), 2013 7th International Conference on*, 2013.

[8] Bishesh Khanal and Désiré Sidibé. Efficient skin detection under severe illumination changes and shadows. In *Proceedings of the 4th international conference on Intelligent Robotics and Applications - Volume Part II*, pages 609–618. Springer-Verlag, Aachen and Germany, 2011.

[9] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors*, 12(2):1437–1454, 2012.

[10] Benjamin Langmann, Klaus Hartmann, and Otmar Loffeld. Depth Camera Technology Comparison and Performance Evaluation. In Pedro Latorre Carmona, J. Salvador Sánchez, and Ana L. N. Fred, editors, *ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, Volume 2, Vilamoura, Algarve, Portugal, 6-8 February, 2012*, pages 438–444. SciTePress, 2012.

[11] Chintan Mishra and Ahmed Khan Zeeshan. *Development and evaluation of a Kinect based Bin-Picking system*. PhD thesis, 2012.

[12] J. Smisek, M. Jancosek, and T. Pajdla, editors. *3D with Kinect: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011.

[13] Todor Stoyanov, Rasoul Mojtahedzadeh, Henrik Andreasson, and Achim J. Lilienthal. Comparative evaluation of range sensor accuracy for indoor mobile robotics and automated logistics applications. *Selected Papers from the 5th European Conference on Mobile Robots (ECMR 2011)*, 61(10):1094–1105, 2013.

[14] M. M. Youssef and V. K. Asari. Human action recognition using hull convexity defect features with multi-modality setups. *Smart Approaches for Human Action Recognition*, 34(15):1971–1979, 2013.