applied sciences

MDPI

Article

Skeleton Tracking Accuracy and Precision Evaluation of Kinect V1, Kinect V2, and the Azure Kinect

Michal Tölgyessy *, Martin Dekan and Ľuboš Chovanec

Institute of Robotics and Cybernetics, Faculty of Electrical Engineering and Information Technology STU in Bratislava, Ilkovičova 3, 812 19 Bratislava, Slovakia; martin.dekan@stuba.sk (M.D.); lubos.chovanec@stuba.sk (Ľ.C.)

* Correspondence: michal.tolgyessy@stuba.sk

Abstract: The Azure Kinect, the successor of Kinect v1 and Kinect v2, is a depth sensor. In this paper we evaluate the skeleton tracking abilities of the new sensor, namely accuracy and precision (repeatability). Firstly, we state the technical features of all three sensors, since we want to put the new Azure Kinect in the context of its previous versions. Then, we present the experimental results of general accuracy and precision obtained by measuring a plate mounted to a robotic manipulator end effector which was moved along the depth axis of each sensor and compare them. In the second experiment, we mounted a human-sized figurine to the end effector and placed it in the same positions as the test plate. Positions were located 400 mm from each other. In each position, we measured relative accuracy and precision (repeatability) of the detected figurine body joints. We compared the results and concluded that the Azure Kinect surpasses its discontinued predecessors, both in accuracy and precision. It is a suitable sensor for human–robot interaction, body-motion analysis, and other gesture-based applications. Our analysis serves as a pilot study for future HMI (human–machine interaction) designs and applications using the new Kinect Azure and puts it in the context of its successful predecessors.

Keywords: depth sensor; Kinect; Azure Kinect; skeleton tracking; skeletal tracking; pose tracking

Citation: Tölgyessy, M.; Dekan, M.; Chovanec, Ľ. Skeleton Tracking Accuracy and Precision Evaluation of Kinect V1, Kinect V2, and the Azure Kinect. *Appl. Sci.* **2021**, *11*, 5756. https://doi.org/10.3390/app11125756

Academic Editor: George Mylonas

Received: 21 May 2021
Accepted: 15 June 2021
Published: 21 June 2021

1. Introduction

The Kinect Xbox 360 has been a revolution in affordable 3D sensing. Initially meant only for the gaming industry, it was soon to be used by scientists, robotics enthusiasts, and hobbyists all around the world. It was later followed by the release of another Kinect—Kinect for Windows. We will refer to the former as Kinect v1 and to the latter as Kinect v2. Both versions have been widely used by the research community for various scientific purposes, such as object detection and object recognition [1–3], mapping and SLAM [4–6], gesture recognition and human–machine interaction (HMI) [7–9], telepresence [10,11], virtual reality, mixed reality, and medicine and rehabilitation [12–16]. However, both sensors are now discontinued and are no longer being officially distributed and sold. In 2019, Microsoft released the Azure Kinect, which is no longer meant for the gaming market in any way; it is promoted as a developer kit with advanced AI sensors for building computer vision and speech models.

Numerous papers focusing on the skeleton tracking capabilities of Kinect v1 and Kinect v2 have been published [17–23]. New ideas making use of the new Azure Kinect have already been published. Albert et al. evaluate the pose tracking performance of the Azure Kinect and Kinect v2 for gait analysis [24]. Lee et al. developed a real-time hand gesture recognition for tabletop holographic display interaction using the Azure Kinect [25]. Manghisi et al. developed a body-tracking-based low-cost solution for monitoring workers' hygiene best practices during pandemics using the Azure Kinect [26]. Lee at al. proposed a robust extrinsic calibration of multiple RGB-D cameras with body tracking and feature matching [27].

Precise 3D body joint detection is crucial for gesture-based and human motion analysis applications. It is especially important in the human–robot interaction field, namely in vision based HRI. This is clearly demonstrated in the Laws of Linear HRI defined in [28]:

1. Every pair of two joints of a human sensed by a robot form a line.
2. Every line defined by first law intersects with the robot's environment in one or two places.
3. Every intersection defined by the second law is a potential navigation target or a potential object of reference for the robot.

For a thorough investigation of the new Azure Kinect, please see [29]. This paper focuses solely on the skeleton tracking capabilities of all Kinect versions. Nevertheless, we performed initial experiments to compare the depth-sensing precision and accuracy of the examined sensors. In the second experiment, we mounted a human-sized figurine to the end effector and placed it in the same positions as the test plate. Positions were located 400 mm from each other. In each position, we measured relative accuracy and precision (repeatability) of the detected figurine body joints.

The main novelty of this paper lies in a skeleton tracking comparison of all three Kinect versions and in using a human-sized figurine precisely positioned by a robotic manipulator with focus on the new Azure Kinect. Evaluation of the new sensor in this regard is very useful for scientists and researchers developing designs and applications where joint tracking without wearable equipment is needed. Since human subjects cannot prevent slight motions while standing still, the plastic figurine is an essential tool to provide reliable precision and accuracy measurements. Skeleton tracking binaries provided by Microsoft are based on the work of Shotton et al. [30], shown in Figure 1.

Figure 1. From left to right—Kinect v1, Kinect v2, Azure Kinect.

The paper is organized as follows. Firstly, we present specifications of all examined sensors. Then, we perform experiments to determine the general precision and accuracy of each sensor. Then, we present experiments where we mounted a plastic figurine on a robotic manipulator to determine the skeletal tracking (body joint detection) precision and accuracy of each sensor. Finally, we determine body joint reliability to show which particular body joints are likely to be detected with lower precision than others. We then summarize our results in the Discussion section.

2. Kinects' Specifications

Both earlier versions of the Kinect have one depth camera and one color camera. The Kinect v1 measures depth with the pattern projection principle, where a known infrared pattern is projected onto the scene and out of its distortion the depth is computed. The Kinect v2 utilizes the continuous wave (CW) intensity modulation approach, which is most commonly used in time-of-flight (ToF) cameras [31].

In a continuous-wave (CW) time-of-flight (ToF) camera, light from an amplitude-modulated light source is backscattered by objects in the camera's field of view, and the

phase delay of the amplitude envelope is measured between the emitted and reflected light. This phase difference is translated into a distance value for each pixel in the imaging array [32].

The Azure Kinect is also based on a CW ToF camera; it uses the image sensor presented in [32]. Unlike Kinect v1 and v2, it supports multiple depth-sensing modes, and the color camera supports a resolution of up to 3840 × 2160 pixels.

The design of the Azure Kinect is shown in Figure 2.

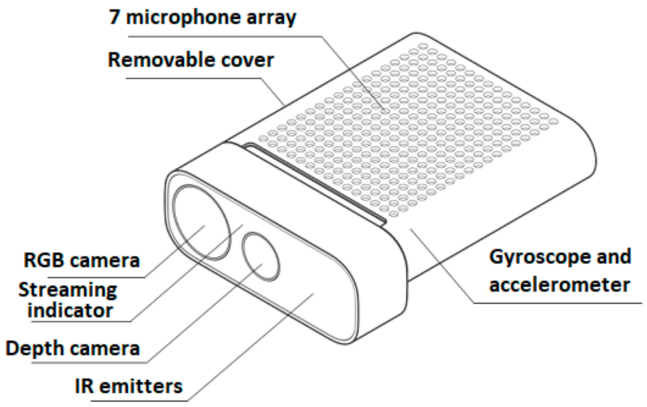


Figure 2. Schematic of the Azure Kinect.

A comparison of the key features of all three Kinects is found in Table 1. All data regarding Azure Kinect is taken from the official online documentation.

Table 1. Comparison of the three Kinect versions [29].

	Kinect v1	Kinect v2	Azure Kinect
Color camera resolution	1280 × 720 px @ 12 fps 640 × 480 px @ 30 fps	1920 × 1080 px @ 30 fps	3840 × 2160 px @30 fps
Depth camera resolution	320 × 240 px @ 30 fps	512 × 424 px @ 30 fps	NFOV unbinned—640 × 576 @ 30 fps NFOV binned—320 × 288 @ 30 fps WFOV unbinned—1024 × 1024 @ 15 fps WFOV binned—512 × 512 @ 30 fps
Depth sensing technology	Structured light–pattern projection	ToF (Time-of-Flight)	ToF (Time-of-Flight)
Field of view (depth image)	57° H, 43° V alt. 58.5° H, 46.6°	70° H, 60° V alt. 70.6° H, 60°	NFOV unbinned—75° × 65° NFOV binned—75° × 65° WFOV unbinned—120° × 120° WFOV binned—120° × 120°
Specified measuring distance	0.4–4 m	0.5–4.5 m	NFOV unbinned—0.5–3.86 m NFOV binned—0.5–5.46 m WFOV unbinned—0.25–2.21 m WFOV binned—0.25–2.88 m
Weight	430 g (without cables and power supply); 750 g (with cables and power supply)	610 g (without cables and power supply); 1390 g (with cables and power supply)	440 g (without cables); 520 g (with cables, power supply is not necessary)

It works in four different modes: NFOV (narrow field-of-view depth mode) unbinned, WFOV (wide field-of-view depth mode) unbinned, NFOV binned, and WFOV binned. The Azure Kinect has both a depth camera and an RGB camera; spatial orientation of the RGB image frame and depth image frame is not identical: there is a 1.3-degree difference. The SDK contains convenience functions for the transformation. These two parts are, according to the SDK, time synchronized by the Azure.

3. Experiments

For general accuracy and precision measurements, we mounted a white reflective plate to the end effector of a robotic manipulator—ABB IRB4600—moved the plate along

the depth axis of each sensor, and performed depth measurements in 7–8 discrete positions (Figure 3), where the distance between adjacent positions was 400 mm. The absolute positioning accuracy of the robot end effector was within 0.02 mm according to the ABB IRB4600 datasheet. Therefore, it was suitable for all our experiments. For all experiments in all positions, we performed 500 measurements which were then used for further statistical evaluation.

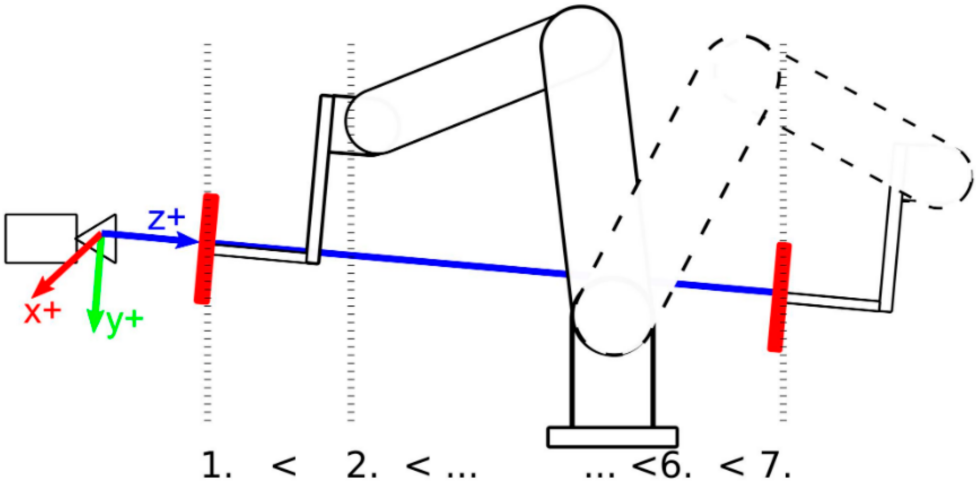


Figure 3. Scheme for general accuracy and precision measurements using robotic manipulator.

We aligned the depth Z axis of each sensor with the robot axis thusly. We aligned the center of the plate with the center of the depth image in the approximate distance of 1500 mm. Then, we moved the plate to an approximate distance of 3000 mm and repeated the procedure; these two points formed a line that corresponded to the depth axis of the depth sensor. We further tested this alignment by moving the plate along the designated line and verified that the depth axis was identified correctly. This procedure is thoroughly covered in [29]. For all experiments and sensors, we used default camera calibrations.

Each sensor was warmed up for one hour, as previous research showed this is necessary to get the best depth results [29].

After aligning the coordinate systems of individual sensors with the coordinate system of the robot, the first experiment we carried out was a comparison of their precision and accuracy. We use these terms in a standard scientific way: by accuracy, we mean how close the measurements are to the true value, and by precision (repeatability), we mean how close the measurements are to each other. As stated before, the alignment did not solve the origins of the individual coordinate systems; therefore, we are not discussing absolute accuracy, but relative accuracy. By relative accuracy, we therefore mean accuracy relative to the first measured position. The core of all our accuracy measurements was measuring points in positions 400 mm away from each other.

3.1. Sensor Precision

The precision of the examined sensors is in Figure 4 in the form of standard deviation. As can be seen, the standard deviation of Kinect v1 considerably grows with distance in both available modes; furthermore, it is the least predictable. This is caused by the sensing technology implemented in Kinect v1, where noise varies for points even in the same distance; Kinect 2 and the Azure Kinect noise have the same character, but the latter has slightly higher precision.

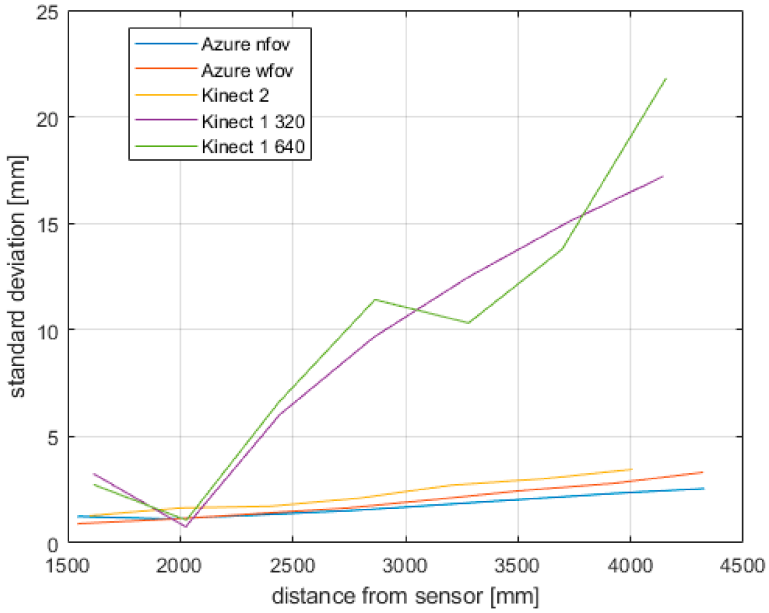


Figure 4. General precision measurements of all three Kinect versions.

3.2. Sensor Accuracy

In Figure 5, the distance variation between measured positions is shown. In an ideal case, the measured variation should be 400 mm between each position since the end effector of the robotic manipulator was programmed to move the test plate in such manner along the depth axis of each sensor. As can be seen in the figure, Kinect v1 had the worst performance, as in the case of precision. For both available resolutions, the difference between the measured and expected value between the first and last position is more than 130 mm. Hence, it is obvious that accuracy of Kinect v1 decreases with growing distance (this behavior is expected due to the projected pattern technology used). As shown in the figure, the position variation of Kinect v2 oscillates around 400 mm. The greatest error was 12 mm but shows little signs of cumulation. The difference between the real and measured position variation of the first and last position is only 10mm. It can be assumed that it would possible to locate a more distant position where the sum of all errors would be 0 mm. The Azure Kinect performed even better; for the NFOV mode, the difference between the real and measured position variation of the first and last position is under 0.5 mm, and for the WFOV mode, it is 1.03 mm. In the WFOV mode, however, up to a particular distance, the Azure Kinect measured higher values, and after that, lower values. Therefore, with growing distance, the accuracy is likely to worsen.

The previous experiment serves as an introduction to the core experiment of this paper; its purpose is to evaluate and compare the skeleton tracking and joint detection of all Kinect versions. In this experiment, the alignment of each sensor was the same as in the former one; this time a fixed human-sized figurine was moved along the depth axis of each sensor (Figure 6). This way, every detected body joint moved 400 mm in between measured positions. As before, we measured relative accuracy, precision, and skeleton detection quality.

In Figure 6C, there is an example of the detected body joints of a person. It is clear that the feet are undetected because the sensor is unable to see them. Each SDK provides the parameters of reliability for all joints, termed the confidence level. We considered a joint to be undetected when the confidence level was lower than the highest level possible. Therefore, even though the figurine could be seen, clothes sometimes caused some joints' confidence level to be lower than the highest level possible, and we considered this case to be an undetected joint.

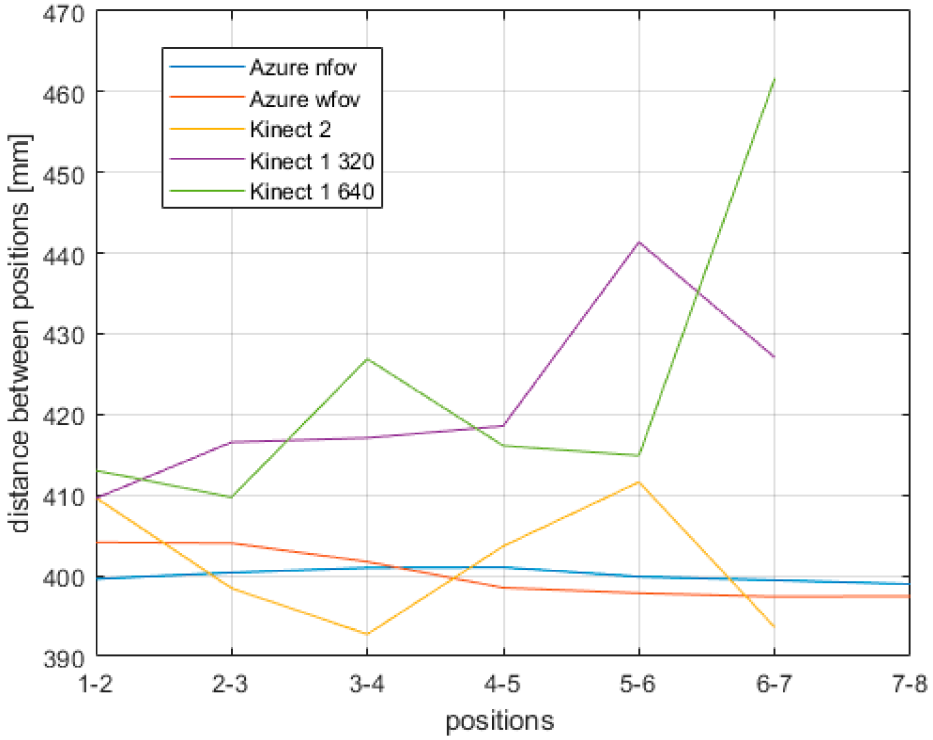
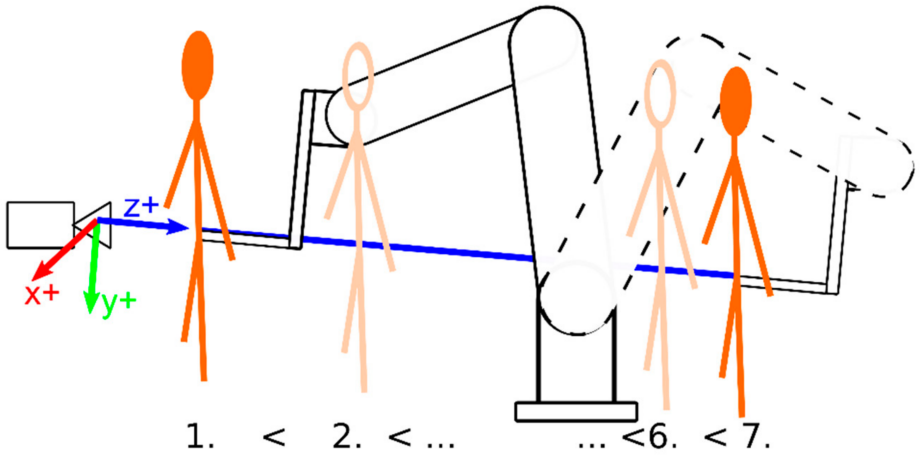
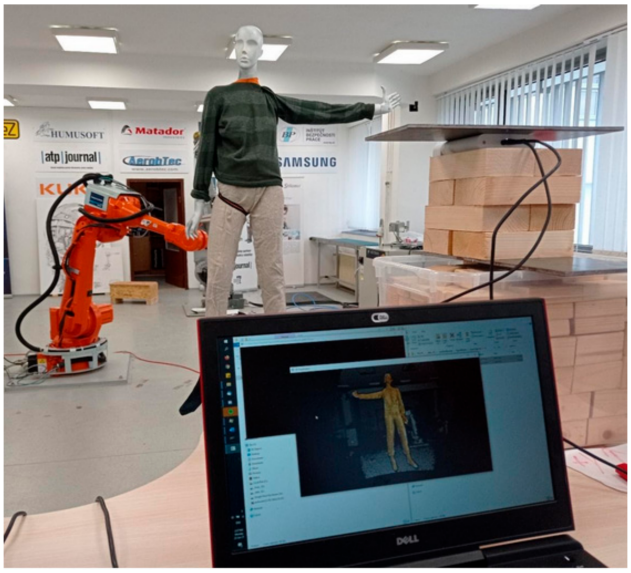


Figure 5. General accuracy measurements of all three Kinect versions.

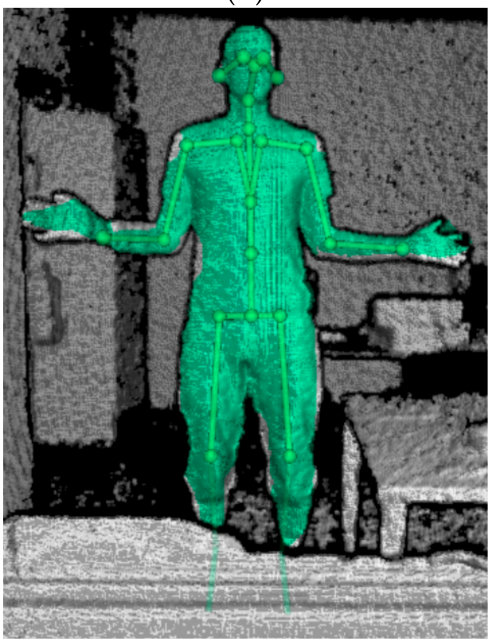


(**A**)

Figure 6. *Cont*.



(**B**)



(**C**)

Figure 6. (**A**) Plastic figurine moved along the Z axis of the examined sensor. (**B**) Picture of the actual laboratory experiment. (**C**) Example of detected and undetected body joints.

We evaluate the skeleton quality as the percentage of undetected joints for a particular sensor and scanning position. The numbers in brackets represent values for joints common for all sensors, as the Azure Kinect gives 32 distinct joints (12 more than Kinect 1). Furthermore, it must be noted that each sensor probably has slightly different metrics for evaluating the reliability of a detected joint (different binary libraries and SDKs); therefore, the presented comparison is only approximate. The result is presented in Table 2.

Table 2. Percentage of undetected joints for each sensor and position.

	Kinect 1	Kinect 1 No Smooth	Kinect 2	Azure NFOV	Azure WFOV
Position 1	15.85	25	0	6.18 (1.09)	0
Position 2	0	0	0	5.63 (10)	6.52 (0.43)
Position 3	30	0	0	18.75 (10)	19.01 (10)
Position 4	30	0	0	18.75 (10)	37.32 (28.13)
Position 5	29.98	15.55	0	18.75 (10)	32.79 (11.23)
Position 6	29.96	30	0.1	18.75 (10)	29.78 (11.3)
Position 7	26.65	35.03	0	18.75 (10)	54.43 (39.22)
Position 8				21.84 (12.1)	41.41 (21.11)

Concerning outage character, most of the time data loss occurred in the form of missing a particular joint. Rarely did a joint drop out back and forth in the same measurement set. Kinect v2 had the least joint outages throughout the tested range; however, it must be noted that its algorithm detects 7 joints less than the Azure Kinect. Most failures happened in the WFOV mode of the Azure Kinect.

Next, we evaluate the precision (repeatability) and accuracy of the skeleton tracking. For this we use only reliable skeleton data: those which the skeleton tracking library classified as most accurate (the tracking SDKs offer several levels of tracking accuracy for each body joint). Thus, it could happen that for a certain position, there are missing data for certain joints; therefore, the final count might vary.

3.3. Skeleton Tracking Precision

The list of all body joints used throughout the rest of the paper is found in Figure 7. In Figures 8–12, there is the precision for particular joints and positions.

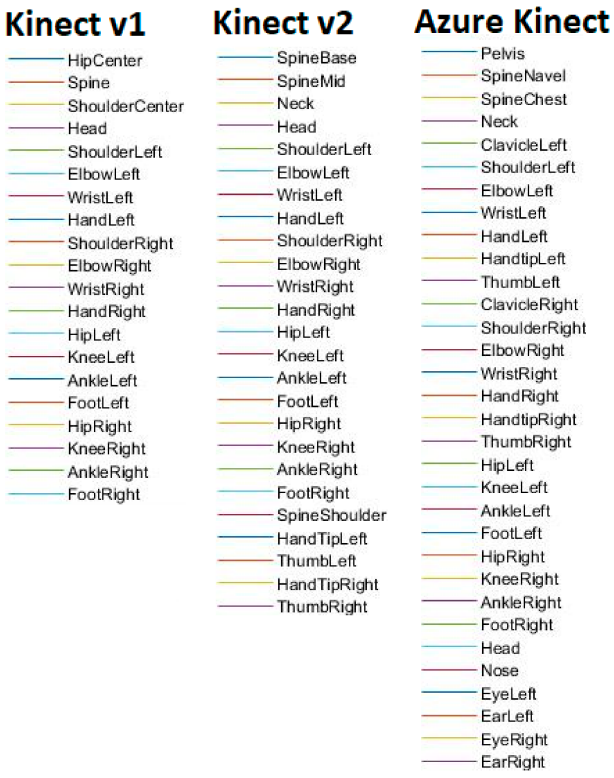


Figure 7. List of skeleton body joints detected by corresponding Kinect sensor (the presented colors are used in all figures throughout the paper).

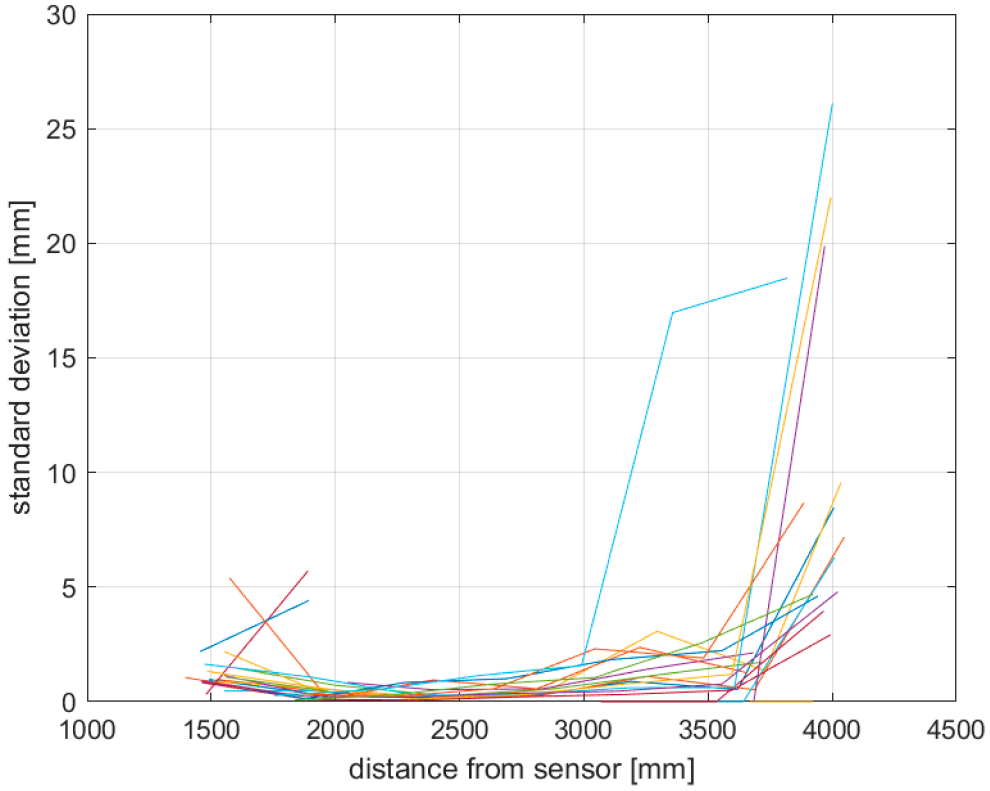


Figure 8. Body joint precision measurements of Kinect v1.

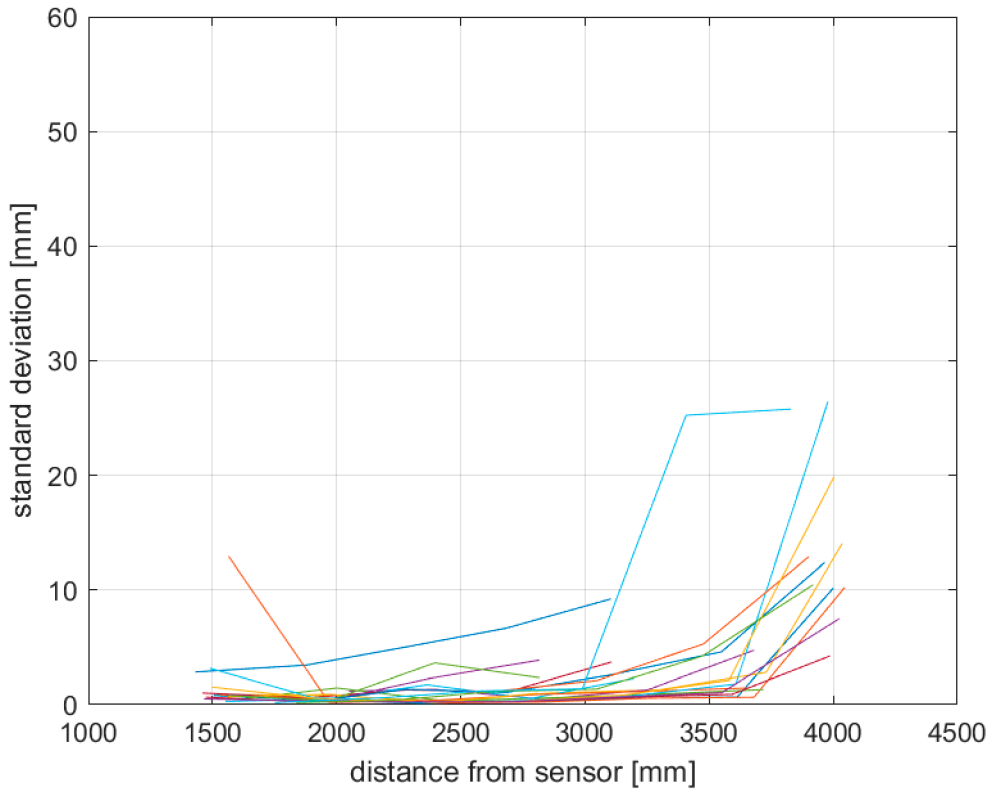


Figure 9. Body joint precision measurements of Kinect v1 with skeleton smoothing turned off.

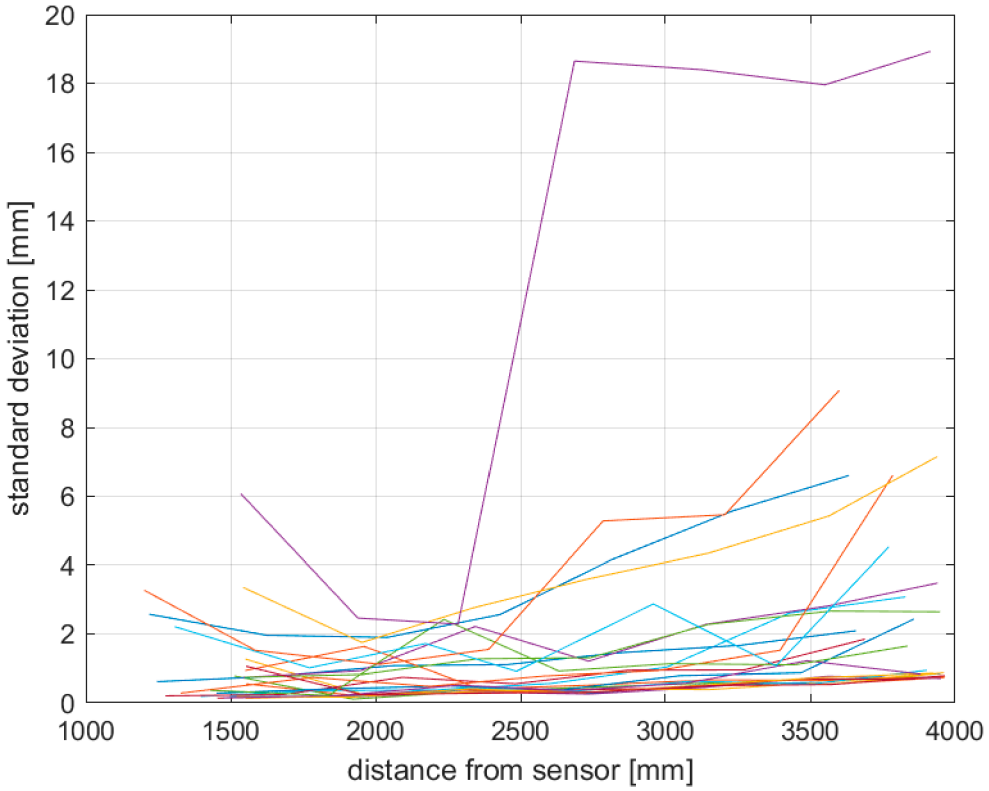


Figure 10. Body joint precision measurements of Kinect v2.

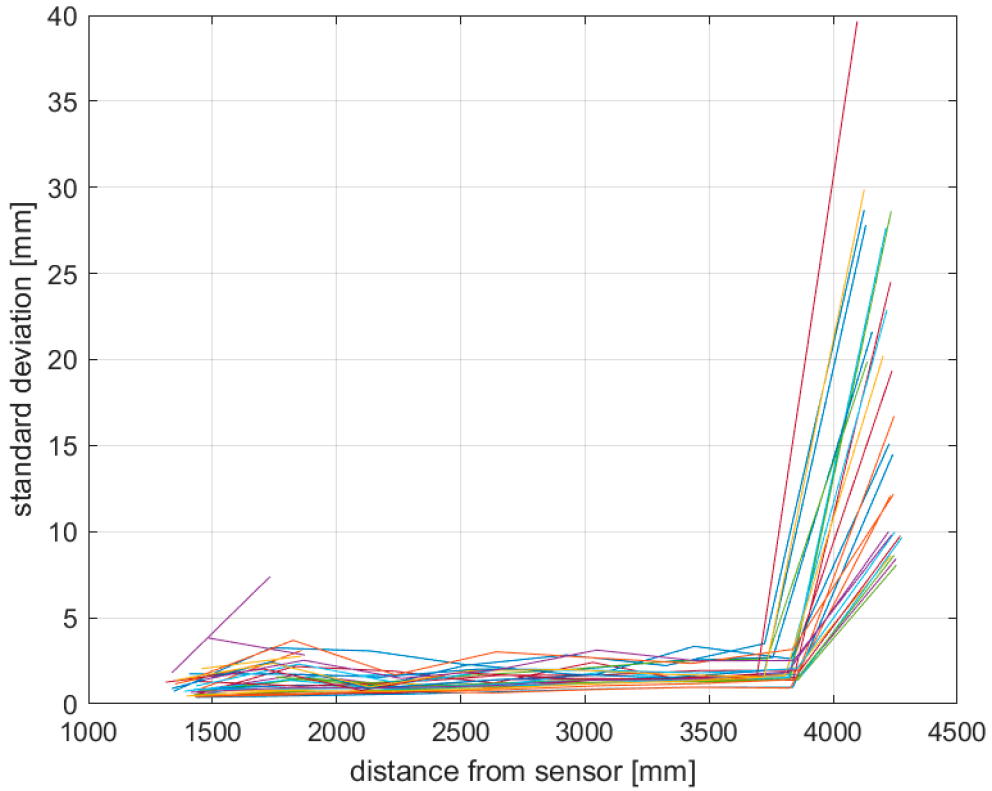


Figure 11. Body joint precision measurements of Azure Kinect in the NFOV mode.

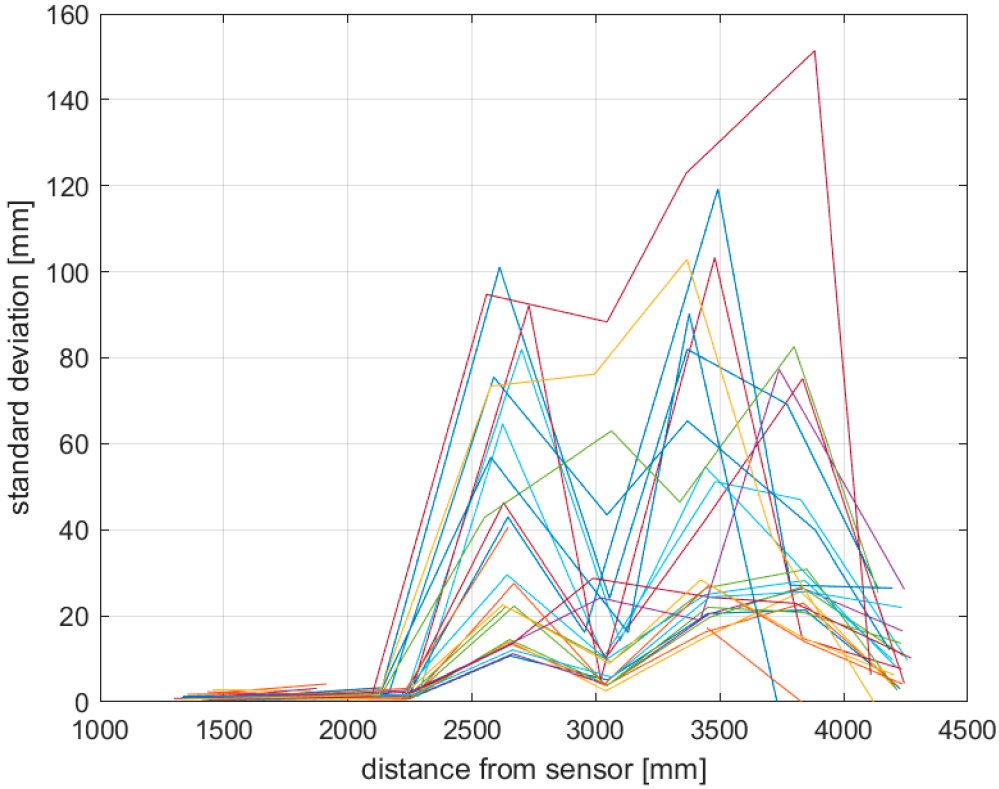


Figure 12. Body joint precision measurements of Azure Kinect in the WFOV mode.

As can be seen, for certain body joints, the standard deviation varies considerably, even within one sensor datum. However, there is a correlation with the measured distance (Figure 4). Furthermore, the Azure Kinect in WFOV mode is quite unusable with distances higher than 2.5 m; this correlates with the amount of unreliable body joint data. In the NFOV mode, the data are stable and reliable up to a 3.6 m distance.

3.4. Skeleton Tracking Accuracy

In this section, we evaluate the accuracy of skeletal tracking. Figures 13–17 show distance differences between measured positions. As with the general accuracy test, Kinect v1 outputs higher position distance variations for detected body joints than the expected 400 mm. The accuracy of Kinect v2 oscillates between less and more than 400 mm. The output of Azure's NFOV mode oscillates more randomly compared to that of Kinect v2 with no correlation to the distance. The WFOV mode of Azure Kinect has considerably worse output when the distance is 2.4 m and higher.

The information in previous figures depicts the variation for each sensor itself, but it is hard to compare the output of individual sensors with each other. Therefore, next, we visualize the average distance the variation of all sensors for each position. The result is depicted in Figure 18. As can be seen, the NFOV mode of the Azure Kinect gives the best output. Its WFOV mode is comparable with Kinect v2; however, from previous figures it is clear that Kinect v2 gives more stable output for particular body joints.

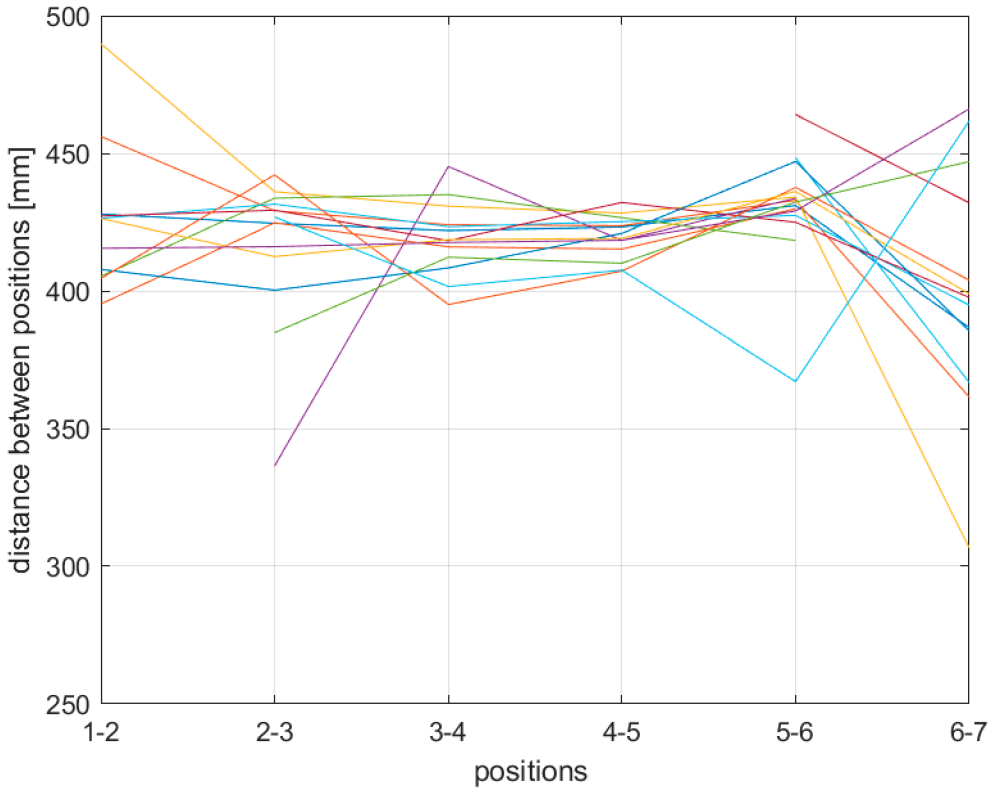


Figure 13. Body joint relative accuracy measurements of Kinect v1.

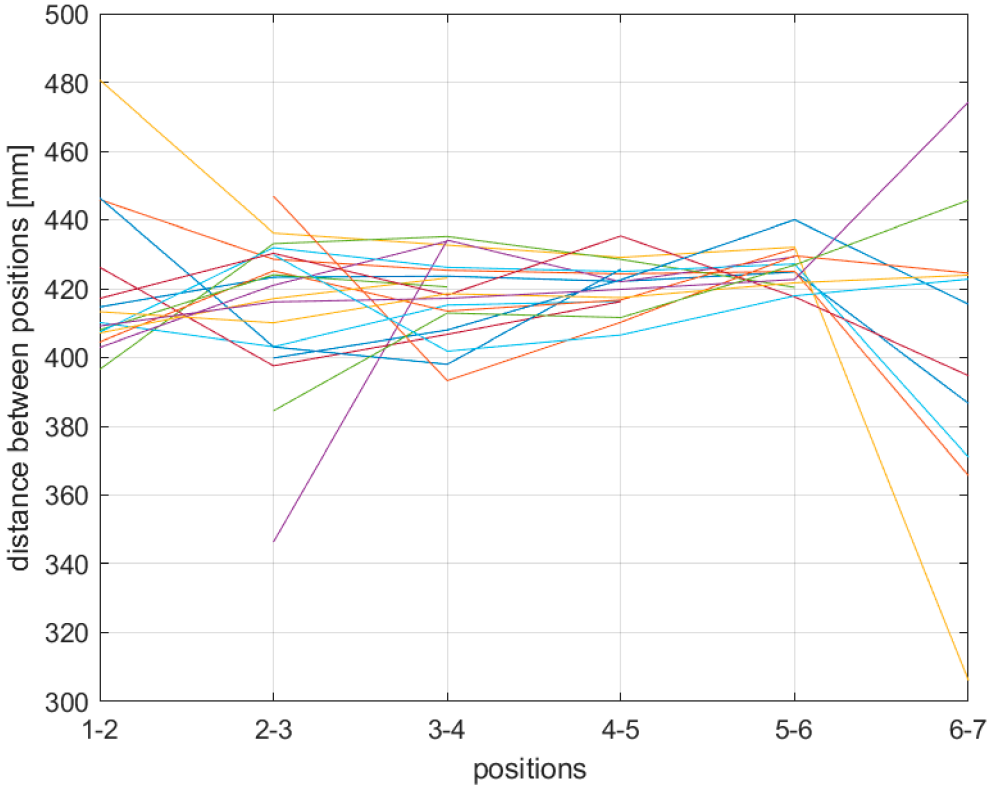


Figure 14. Body joint relative accuracy measurements of Kinect v1 with skeleton smoothing turned off.

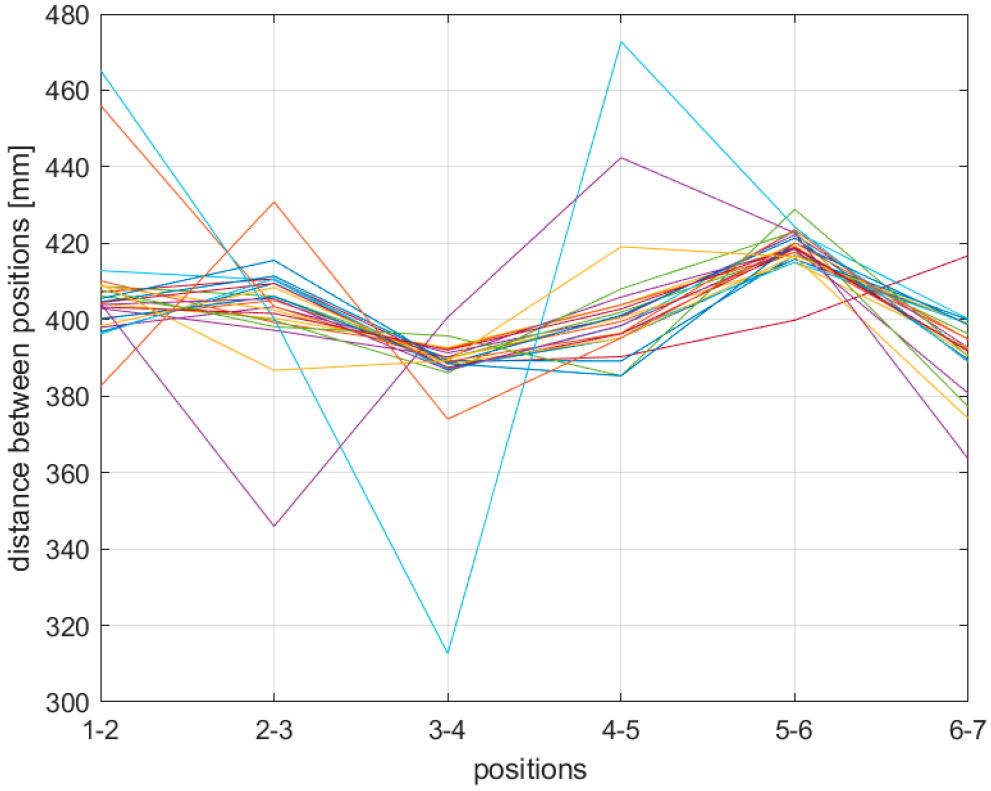


Figure 15. Body joint relative accuracy measurements of Kinect v2.

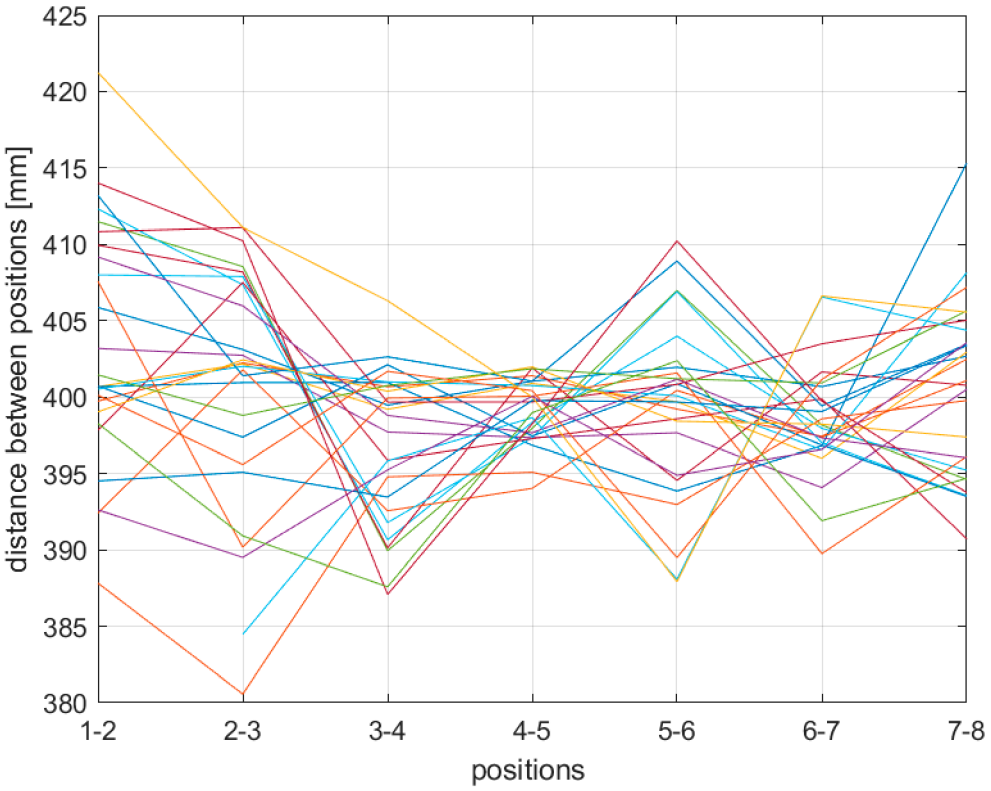


Figure 16. Body joint relative measurements of Azure Kinect in the NFOV mode.

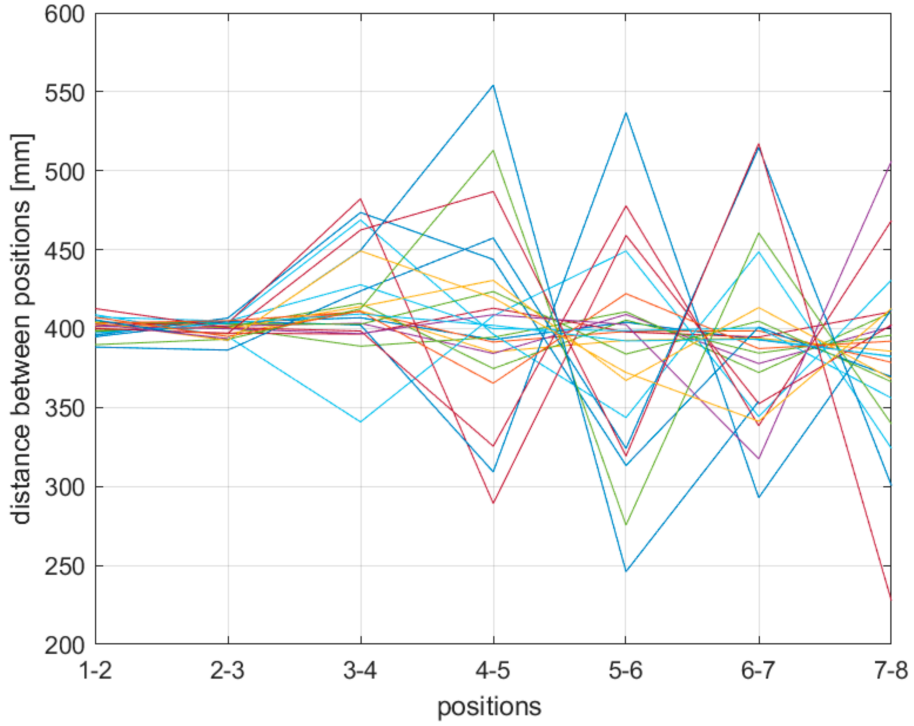


Figure 17. Body joint relative measurements of Azure Kinect in the WFOV mode.

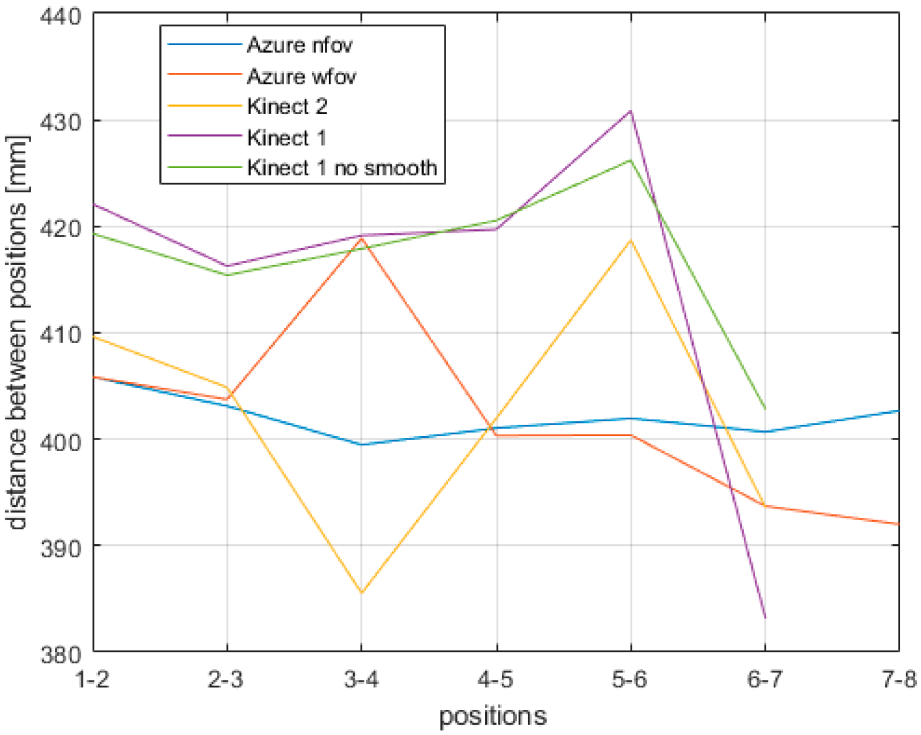


Figure 18. Body joint average position distance variation of all sensors (only body joints common for all sensors were are included).

4. Body Joint Reliability

Next, we identify the body joints that are the least reliable: those whose noise is the greatest, and who have the most frequent outages. We consider an outage of a joint to be when the skeleton detection binary removes it from the highest detection reliability class.

In Figures 19–23, there are figurine body joint data from position 5. The size of each ellipsoid represents the standard deviation of a particular joint 3D position multiplied by 10 for better visual clarity. Data for other positions had a similar character, but for the sake of space, we present only the representative sample.

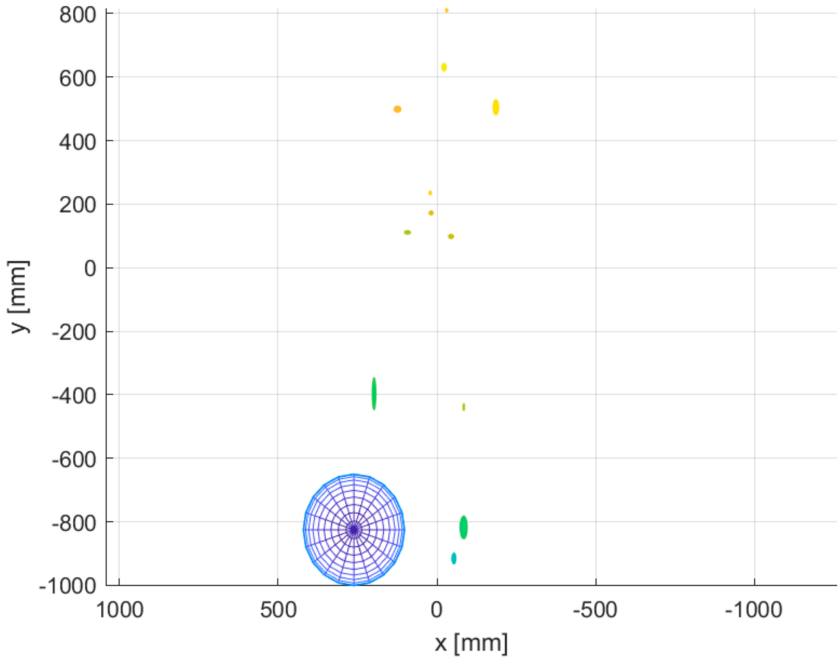


Figure 19. Body joint noise of Kinect v1.

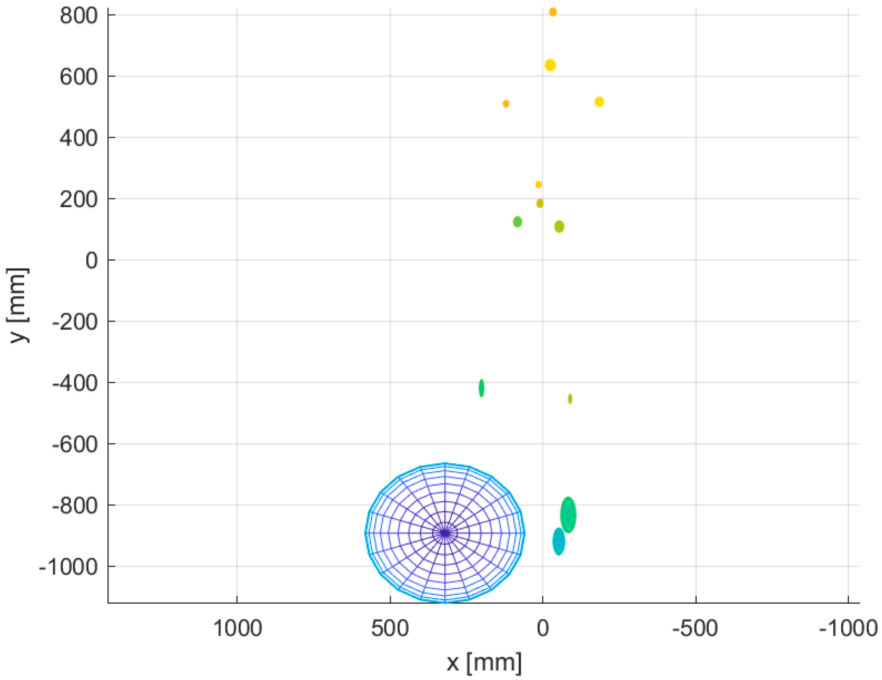


Figure 20. Body joint noise of Kinect v1 with skeleton smoothing turned off.

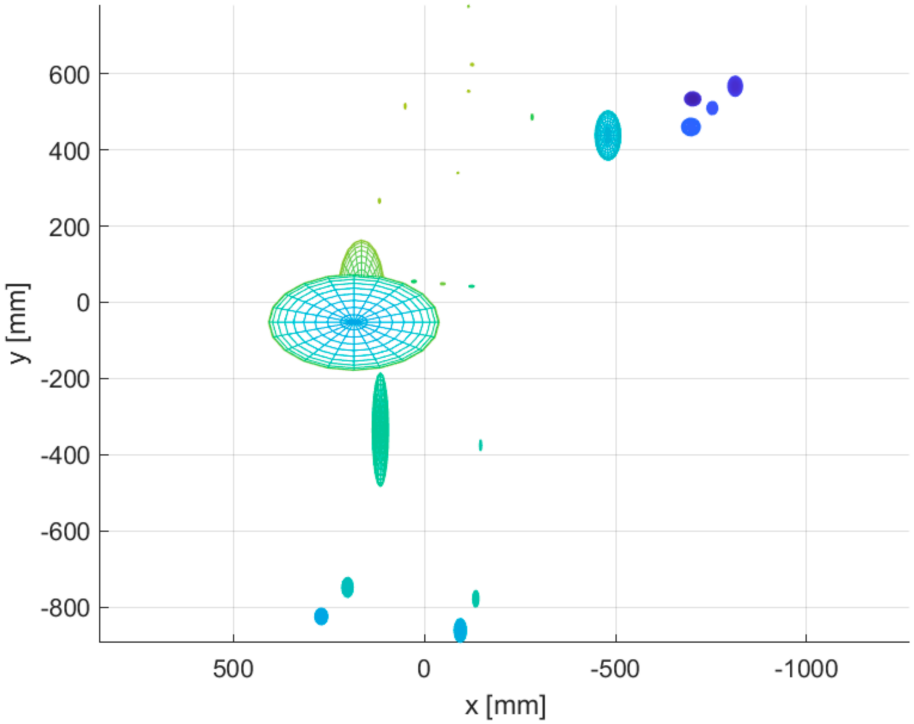


Figure 21. Body joint noise of Kinect v2.

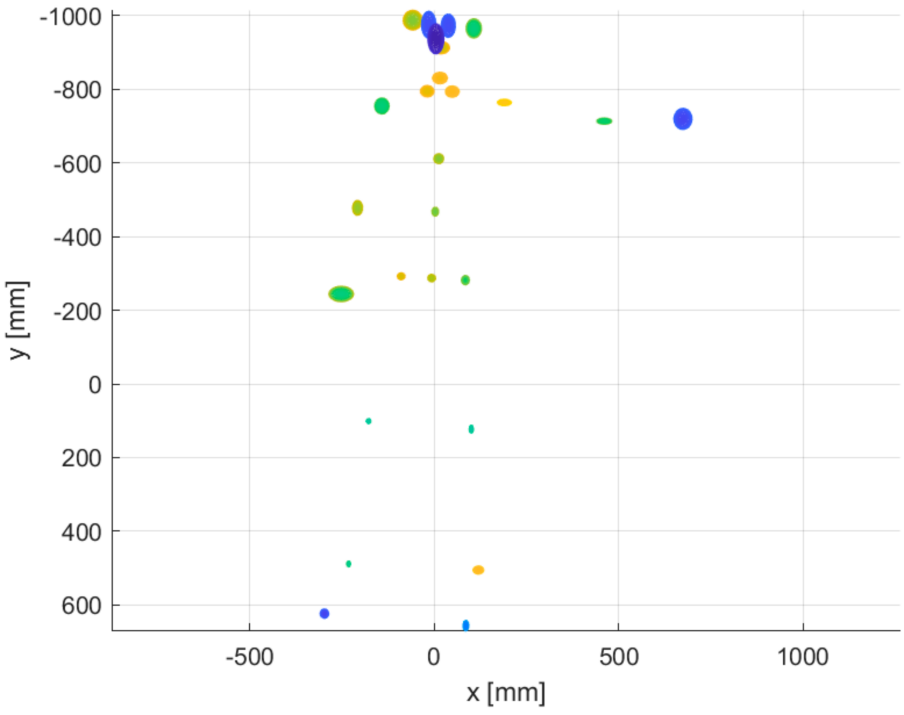


Figure 22. Body joint noise of Azure Kinect in the NFOV mode.

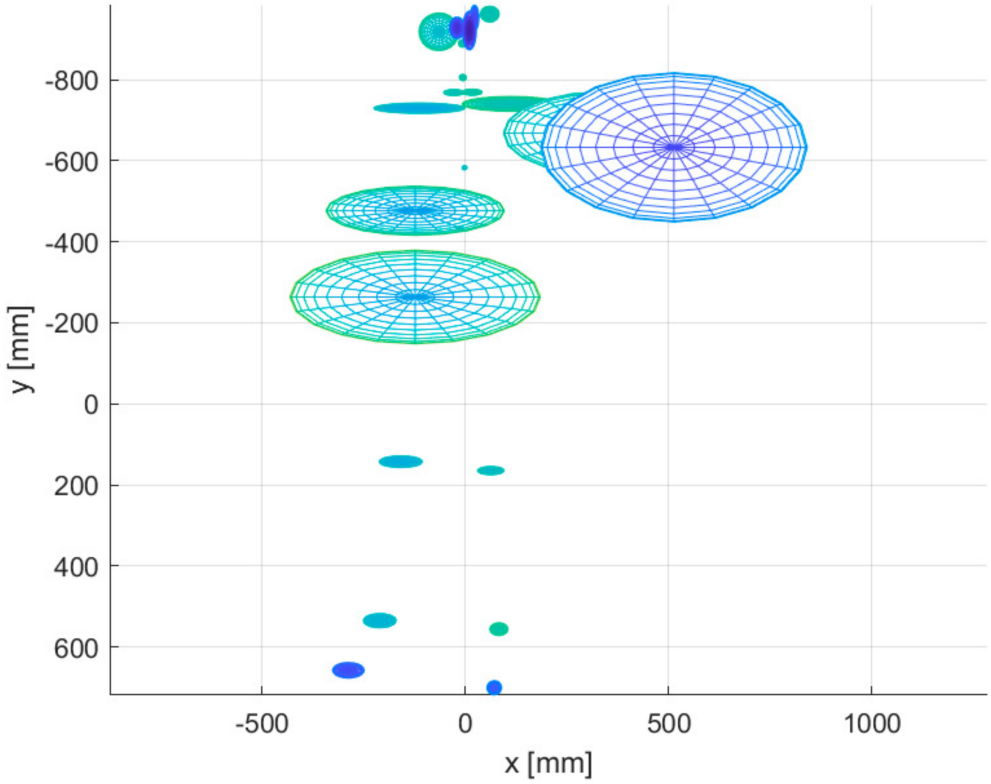


Figure 23. Body joint noise of Azure Kinect in the WFOV mode.

It is clear from the acquired data that the greatest instability occurs at the ends of limbs. In the case of the Azure Kinect, there is higher noise in the head location, but it is caused by the fact that the head is represented by more than one joint in the Azure skeletal tracking detection system. The chest is the most stable segment of the skeleton tracking output for all Kinect versions. It can be assumed that if a body joint is followed by another successfully located joint, its position is determined more precisely.

The previous figures show data for one typical position only to illustrate the general behavior of skeleton tracking. We include the two following tables which contain only key elements of sensor behavior for all positions to quantify the most important details of all measurements.

In Table 3, joint distances between individual positions where the expected result is 400 mm are shown. For each sensor and the respective mode, there are three rows. The first row is the average distance change of all joints; the second row is the minimal average distance change of a joint (the one with the lowest value); and the third row is the maximal average distance change of a joint (the one with the highest value). As can be seen, the Azure Kinect values are closest to 400 mm; Kinect v2 has similar values, but the maximal values are considerably higher compared to the Azure.

In Table 4, the noise values expressed by standard deviation are shown. As in Table 3, we present the average noise of all joints and the maximal and minimal noise for each respective position.

Table 3. Key data for all positions and sensors.

		Pos 1–2	Pos 2–3	Pos 3–4	Pos 4–5	Pos 5–6	Pos 7–8	
Azure NFOV	Avg dst	405.8667	403.1197	399.4731	401.0482	401.9613	400.6904	402.6862
	Min dst	390.037	382.7666	389.2931	396.2346	390.295	391.9491	392.9188
	Max dst	416.2143	413.291	404.8484	404.1863	412.4341	408.7557	417.5213
Azure WFOV	Avg dst	405.8589	403.735	418.8697	400.3382	400.39	393.7127	392.0062
	Min dst	391.5172	389.5792	344.0758	292.5254	278.7621	296.1326	303.7741
	Max dst	440.8246	409.8515	485.4536	516.0719	539.9829	517.9588	509.3196
Kinect 2	Avg dst	409.6489	404.8747	385.5388	401.9627	418.7103	393.5865	0
	Min dst	396.249	397.1817	312.6303	385.3205	399.8347	377.1883	0
	Max dst	465.2497	411.3906	395.7215	472.7374	428.8178	416.779	0
Kin 1	Avg dst	422.1129	416.271	419.1457	419.7023	430.8499	383.1812	0
	Min dst	395.228	336.4606	395.0333	407.2178	367.0855	253.103	0
	Max dst	489.8325	442.156	445.1825	432.0966	464.095	466.0918	0
Kin 1 no smooth	Avg dst	419.3267	415.4121	417.8968	420.5548	426.2065	402.8094	0
	Min dst	396.5619	346.3239	393.2796	406.5002	417.6078	306.1316	0
	Max dst	480.6929	446.8681	435.2085	435.3634	440.0262	474.2191	0

Table 4. Key noise data for all positions and sensors.

		Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 7	Pos 8	Pos 9
Azure NFOV	Avg std	0.7955	1.4593	1.1816	1.2902	1.5918	1.6813	1.8967	17.5621
	Min std	0.3868	0.4701	0.5686	0.6996	0.8603	0.9727	0.96	8.4145
	Max std	1.7289	3.2511	3.0636	2.245	2.832	3.3347	3.499	28.6737
Azure WFOV	Avg std	0.9367	1.3618	1.4429	40.1682	16.0603	44.1544	37.2374	12.419
	Min std	0.382	0.6155	0.726	10.6573	2.5789	14.7353	13.8637	2.7271
	Max std	2.1455	4.1929	3.3282	101.0889	62.9689	119.2252	82.6349	26.4784
Kinect 2	Avg std	0.5162	0.452	0.7477	0.5846	0.9744	1.145	1.8635	0
	Min std	0.1336	0.1087	0.2787	0.2575	0.3879	0.5417	0.6947	0
	Max std	2.2119	1.6368	2.4234	1.3069	2.8729	2.8123	6.6177	0
Kin 1	Avg std	1.3966	0.8655	0.3485	0.5298	1.18	1.9019	10.1528	0
	Min std	0.319	0.0781	0.0704	0.2106	0	0	0	0
	Max std	5.3957	5.7061	0.9518	1.1491	3.0647	16.9649	26.1015	0
Kin 1 no smooth	Avg std	1.8913	0.6573	1.0228	1.2459	1.8447	4.0605	15.4636	0
	Min std	0.2877	0.1858	0.1416	0.234	0.5442	0.6522	4.2336	0
	Max std	12.9359	3.437	5.0265	6.6438	9.2239	25.2314	51.0188	0

5. Discussion

In this paper, we presented the experimental results of the general accuracy and precision of all Kinect versions obtained by measuring a plate mounted to a robotic manipulator end effector which was moved along the depth axis of each sensor. In the second experiment, we mounted a human-sized figurine to the end effector and placed it in the same positions as the test plate. In each position, we measured the relative accuracy and precision (repeatability) of the detected figurine body joints.

Based on the presented data, it is safe to say that the Azure Kinect in NFOV mode is more accurate and precise in skeleton tracking than Kinect v1. Kinect v2 standard deviation values are very similar to those of the Azure Kinect; however, as is clear from Figure 18, the NFOV mode of the Azure Kinect is much more stable in terms of accuracy than Kinect v2. Furthermore, the range of the Azure Kinect is higher; it was the only sensor that could capture the figurine even behind the manipulator range; plus, the WFOV mode has a wider

angular resolution than Kinect v2. Therefore, since previous Kinect versions have been discontinued, the new sensor can certainly stand in their place. Furthermore, the Azure detects more body joints and works on the Linux platform (unlike previous versions), which is prevalent, especially in robotics. Its WFOV mode is imprecise, but its wide field of view could be used for simple people detection applications.

It is a hindrance of this study that the skeleton detection binaries were distributed only with each specific Kinect version, but based on our observations, the versions for Kinect v1 and v2 behave similarly. Namely, one often must perform motion towards the sensor for her or his body joints to be reliably tracked. This caused some problems while we were performing experiments, as we were forced to move the figurine back and forth before stabilizing it in its current measuring position. The Azure Kinect detects the skeleton much more smoothly, and no movement towards the sensor is necessary. Other hindrances include the reflectivity of the figurine. We had it clothed, including socks, and painted with a reflective paint, but it is possible that palm and face detection were affected by this. We were unable to align the origins of the Kinect frames; therefore, we could only measure relative accuracy.

Our future work will be focused on advancing the field of HRI (human–robot interaction). We are working on gesture-based interactive robotic systems and plan to develop applications that make use of the Laws of Linear HRI. Furthermore, we plan to compare the Azure skeleton tracking capabilities with other available systems, such as Intel Real Sense technology, OpenPose, DeepPose, and VNect.

To conclude, the Azure Kinect is a promising small and versatile device with a wide range of uses, ranging from object recognition, object reconstruction, mobile robot mapping, navigation, obstacle avoidance, and SLAM to object tracking, people tracking and detection, HCI (human–computer interaction), HMI (human–machine interaction), HRI (human–robot interaction), gesture recognition, virtual reality, telepresence, medical examination, biometry and more.

6. Patents

This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

Author Contributions: Conceptualization, M.T. and M.D.; methodology, M.D. and Ľ.C.; software, M.T.; validation, M.T., M.D. and Ľ.C.; formal analysis, M.D.; investigation, M.T.; resources, M.T. and Ľ.C.; data curation, M.D.; writing—original draft preparation, M.T.; writing—review and editing, M.T.; visualization, M.D.; supervision, M.T.; project administration, M.T.; funding acquisition, Peter Hubinský, Martin Dekan. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by APVV-17-0214, VEGA 1/0775/20 and VEGA 1/0754/19.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elaraby, A.F.; Hamdy, A.; Rehan, M. A Kinect-Based 3D Object Detection and Recognition System with Enhanced Depth Estimation Algorithm. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 247–252. [CrossRef]
2. Tanabe, R.; Cao, M.; Murao, T.; Hashimoto, H. Vision based object recognition of mobile robot with Kinect 3D sensor in indoor environment. In Proceedings of the 2012 Proceedings of SICE Annual Conference (SICE); Akita, Japan, 20–23 August 2012; pp. 2203–2206.
3. Manap, M.S.A.; Sahak, R.; Zabidi, A.; Yassin, I.; Tahir, N.M. Object Detection using Depth Information from Kinect Sensor. In Proceedings of the 2015 IEEE 11th International Colloquium on Signal Processing & Its Applications (CSPA), Kuala Lumpur, Malaysia, 6–8 March 2015; pp. 160–163.

4. Xin, G.X.; Zhang, X.T.; Wang, X.; Song, J. A RGBD SLAM algorithm combining ORB with PROSAC for indoor mobile robot. In Proceedings of the 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, China, 19–20 December 2015; pp. 71–74.
5. 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. *Int. J. Robot. Res.* **2012**, *31*, 647–663. [CrossRef]
6. Ibragimov, I.Z.; Afanasyev, I.M. Comparison of ROS-based visualslam methods in homogeneous indoor environment. In Proceedings of the 2017 14th Workshop on Positioning, Navigation and Communications (WPNC), Bremen, Germany, 25–26 October 2017.
7. Plouffe, G.; Cretu, A. Static and dynamic hand gesture recognition in depthdata using dynamic time warping. *IEEE Trans. Instrum. Meas.* **2016**, *65*, 305–316. [CrossRef]
8. Wang, C.; Liu, Z.; Chan, S. Superpixel-based hand gesture recognition withKinect depth camera. *IEEE Trans. Multimed.* **2015**, *17*, 29–39. [CrossRef]
9. Ren, Z.; Yuan, J.; Meng, J.; Zhang, Z. Robust part-based hand gesture recognition using kinect sensor. *IEEE Trans. Multimed.* **2013**, *15*, 1110–1120. [CrossRef]
10. Avalos, J.; Cortez, S.; Vasquez, K.; Murray, V.; Ramos, O.E. Telepres-ence using the kinect sensor and the nao robot. In Proceedings of the 2016 IEEE 7th Latin American Symposium on Circuits & Systems (LASCAS), Florianopolis, Brazil, 28 February–2 March 2016; pp. 303–306.
11. Berri, R.; Wolf, D.; Osório, F.S. Telepresence Robot with Image-Based Face Tracking and 3D Perception with Human Gesture Interface Using Kinect Sensor. In Proceedings of the 2014 Joint Conference on Robotics: SBR-LARS Robotics Symposium and Robocontrol, Sao Carlos, Brazil, 18–23 October 2014; pp. 205–210.
12. Tao, G.; Archambault, P.S.; Levin, M.F. Evaluation of Kinect skeletal tracking in a virtual reality rehabilitation system for upper limb hemiparesis. In Proceedings of the 2013 International Conference on Virtual Rehabilitation (ICVR), Philadelphia, PA, USA, 26–29 August 2013; pp. 164–165.
13. Satyavolu, S.; Bruder, G.; Willemsen, P.; Steinicke, F. Analysis of IR-based virtual reality tracking using multiple Kinects. In Proceedings of the 2012 IEEE Virtual Reality (VR), Costa Mesa, CA, USA, 4–8 March 2012; pp. 149–150.
14. Gotsis, M.; Tasse, A.; Swider, M.; Lympouridis, V.; Poulos, I.C.; Thin, A.G.; Turpin, D.; Tucker, D.; Jordan-Marsh, M. Mixed realitygame prototypes for upper body exercise and rehabilitation. In Proceedings of the 2012 IEEE Virtual Reality Workshops (VRW), Costa Mesa, CA, USA, 4–8 March 2012; pp. 181–182.
15. Heimann-Steinert, A.; Sattler, I.; Otte, K.; Röhling, H.M.; Mansow-Model, S.; Müller-Werdan, U. Using New Camera-Based Technologies for Gait Analysis in Older Adults in Comparison to the Established GAITRite System. *Sensors* **2019**, *20*, 125. [CrossRef] [PubMed]
16. Volák, J.; Koniar, D.; Hargas, L.; Jablončík, F.; Sekel'Ova, N.; Durdík, P. RGB-D imaging used for OSAS diagnostics. In Proceedings of the 2018 ELEKTRO, Mikulov, Czech Republic, 21–23 May 2018; pp. 1–5. [CrossRef]
17. Zhu, H.; Pun, C. Human action recognition with skeletal information from depth camera. In Proceedings of the 2013 IEEE International Conference on Information and Automation (ICIA), Yinchuan, China, 26–28 August 2013; pp. 1082–1085. [CrossRef]
18. Wei, T.; Lee, B.; Qiao, Y.; Kitsikidis, A.; Dimitropoulos, K.; Grammalidis, N. Experimental study of skeleton tracking abilities from microsoft kinect non-frontal views. In Proceedings of the 2015 3DTV-Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), Lisbon, Portugal, 8–10 July 2015; pp. 1–4. [CrossRef]
19. Chen, N.; Chang, Y.; Liu, H.; Huang, L.; Zhang, H. Human Pose Recognition Based on Skeleton Fusion from Multiple Kinects. In Proceedings of the 2018 37th Chinese Control Conference (CCC), Wuhan, China, 25–27 July 2018; pp. 5228–5232. [CrossRef]
20. Gündüz, A.F.; Şen, M.O.; Karci, A.; Yeroğlu, C. Artificial immune system optimization based duplex kinect skeleton fusion. In Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 16–17 September 2017; pp. 1–5. [CrossRef]
21. Cao, M.; Hashimoto, H. Specific person recognition and tracking of mobile robot with Kinect 3D sensor. In Proceedings of the IECON 2013–39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, Austria, 10–13 November 2013; pp. 8323–8328. [CrossRef]
22. Chen, J.; Wu, X.; Guo, T. 3-D real-time image matching based on kinect skeleton. In Proceedings of the 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, Canada, 4–7 May 2014; pp. 1–4. [CrossRef]
23. Fachri, M.; Hudhajanto, R.P.; Mulyadi, I.H. Wayang Kulit Movement Control System Using Kinect Sensor. In Proceedings of the 2019 2nd International Conference on Applied Engineering (ICAE), Batam, Indonesia, 2–3 October 2019; pp. 1–4. [CrossRef]
24. Albert, J.A.; Owolabi, V.; Gebel, A.; Brahms, C.M.; Granacher, U.; Arnrich, B. Evaluation of the Pose Tracking Performance of the Azure Kinect and Kinect v2 for Gait Analysis in Comparison with a Gold Standard: A Pilot Study. *Sensors* **2020**, *20*, 5104. [CrossRef] [PubMed]
25. Lee, C.; Kim, J.; Cho, S.; Kim, J.; Yoo, J.; Kwon, S. Development of Real-Time Hand Gesture Recognition for Tabletop Holographic Display Interaction Using Azure Kinect. *Sensors* **2020**, *20*, 4566. [CrossRef] [PubMed]
26. Manghisi, V.M.; Fiorentino, M.; Boccaccio, A.; Gattullo, M.; Cascella, G.L.; Toschi, N.; Pietroiusti, A.; Uva, A.E. A Body Tracking-Based Low-Cost Solution for Monitoring Workers' Hygiene Best Practices during Pandemics. *Sensors* **2020**, *20*, 6149. [CrossRef] [PubMed]

27. Lee, S.-H.; Yoo, J.; Park, M.; Kim, J.; Kwon, S. Robust Extrinsic Calibration of Multiple RGB-D Cameras with Body Tracking and Feature Matching. *Sensors* **2021**, *21*, 1013. [CrossRef] [PubMed]
28. Tölgyessy, M.; Dekan, M.; Duchoň, F.; Rodina, J.; Hubinský, P.; Chovanec, L. Foundations of Visual Linear Human–Robot Interaction via Pointing Gesture Navigation. *Int. J. Soc. Robot.* **2017**, *9*, 509–523. [CrossRef]
29. Tölgyessy, M.; Dekan, M.; Chovanec, Ľ.; Hubinský, P. Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2. *Sensors* **2021**, *21*, 413. [CrossRef] [PubMed]
30. Shotton, J.; Girshick, R.; Fitzgibbon, A.; Sharp, T.; Cook, M.; Finocchio, M.; Moore, R.; Kohli, P.; Criminisi, A.; Kipman, A.; et al. Efficient human pose estimation from single depth images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2821–2840. [CrossRef] [PubMed]
31. Sarbolandi, H.; Lefloch, D.; Kolb, A. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Comput. Vis. Image Underst.* **2015**, *139*, 1–20. [CrossRef]
32. Bamji, C.S.; Mehta, S.; Thompson, B.; Elkhatib, T.; Wurster, S.; Akkaya, O.; Payne, A.; Godbaz, J.; Fenton, M.; Rajasekaran, V.; et al. IMpixel 65nm BSI 320MHz demodulated TOF Image sensor with 3 µm global shutter pixels and analog binning. In Proceedings of the 2018 IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, CA, USA, 11–15 February 2018; pp. 94–96.