Depth Sensor Selection for Specific Application

1st Jakub Bajzik
Dep. of Mechatronics and Electronics
University of Zilina
Zilina, Slovakia
jakub.bajzik@fel.uniza.sk

2nd Dusan Koniar
Dep. of Mechatronics and Electronics
University of Zilina
Zilina, Slovakia
dusan.koniar@fel.uniza.sk

3rd Libor Hargas
Dep. of Mechatronics and Electronics
University of Zilina
Zilina, Slovakia
libor.hargas@fel.uniza.sk

4th Jozef Volak
Dep. of Mechatronics and Electronics
University of Zilina
Zilina, Slovakia
jozef.volak@fel.uniza.sk

5th Silvia Janisova
Dep. of Mechatronics and Electronics
University of Zilina
Zilina, Slovakia
silvia.janisova@fel.uniza.sk

***Abstract*—To find the solutions to various technical issues in image processing we often want to be able to obtain not only color but also depth information. Information about the depth helps with a better understanding and description of the scene. As a practical matter, we aim to convert the geometrical information about the object to the digital format with the highest possible accuracy. For such a conversion there are many existing methods of creating a 3D model, e.g. photogrammetry, laser scanners. These methods provide high-quality 3D information; however, they are high-priced and often limited by the size of both scanner and object and the scanning time. As an alternative, we present an optical depth sensor that can be used excluding other emerging disadvantages with lower quality of an output model though. In the future, we consider the usage of the selected sensor for specific biomedical application: diagnostic support of obstructive sleep apnea.**

***Index Terms*—optical depth sensor, 3D model, multi-camera system, error estimation**

I. Introduction

Optical depth sensors convert the 3D information about the screening scene into the 2-dimensional plane, that can be converted by the reversed reconstruction back to 3-dimensional space. A combination of depth image and the color image from the RGB sensor allows us to create a textured model of the scene. Optical depth sensors allow capturing the non-linear nature of craniofacial anatomy needed for prediction of obstructive sleep apnea, such as shape and contour in a faster, cheaper, more readily available way, compared with other imaging techniques [1]. A key attribute for the mentioned application is a geometrical precision of an output model. Therefore, it is important to choose a suitable sensor with the least measurement error.

The goal of multi-view is to reconstruct a complete 3D object model form a collection of images taken from known camera viewpoints. To meet the requirement of a complete model of head and neck without any artifacts we aim to evaluate each of multicamera systems based on: The Intel® RealSense™ Depth Camera D415 Series sensors, Stereolabs ZED Mini depth camera, Microsoft Kinect for Windows V2 and Intel® RealSense™ Camera SR300 and offer a comparison of individual operate technologies.

II. Theory

A. Time of flight camera measurement model

ToF cameras are optical sensors that provide information about the depth of the scene. They contain an active light source that generates an amplitude modulated signal. The signal may have a continuous or impulse character. Most of ToF cameras emit amplitude modulated continuous wave (AMCW) with a frequency near IR for illuminating the scene [2]. Depth measurement is based on the amplitude measurement of the phase difference of the transmitted and received modulated signal, as shown in Figure 1. The depth information for each pixel can be calculated by the synchronous demodulation of the received modulated light in the detector. The demodulation can be performed by interleaving with the original modulated signal.

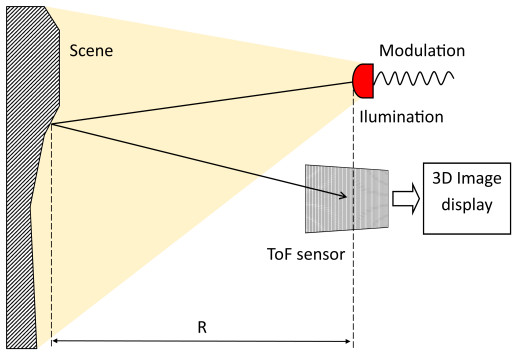


Fig. 1: ToF camera phase-measurement principle.

$$c(\tau) = s(t) \otimes g(t) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} s(t) \cdot g(t+\tau) dt \qquad (1)$$

where $s(t)$ is the received optical signal and the $g(t)$ is transmitted (original) signal. Using specific functions for ToF cameras the equations are:

$$g(t) = \cos \omega t \tag{2}$$

$$s(t) = 1 + a \cdot \cos\left(\omega t - \varphi\right) \tag{3}$$

$$c(\tau) = \varphi_{sg}(\tau) = \frac{a}{2} \cdot cos(\varphi + \omega\tau) \tag{4}$$

where a is modulation amplitude and φ is phase shift. This function is calculated for four different ωt arguments that are shifted from 0 by 90°. The received signal is mostly superimposed on a background image, which requires adding the offset b to the correlation function:

$$C(\tau) = c(\tau) + b \tag{5}$$

$$\begin{aligned} C(\tau_0) &= c(\tau_0) + b = \frac{a}{2} \cdot cos(\varphi) + b \\ C(\tau_1) &= c(\tau_1) + b = -\frac{a}{2} \cdot sin(\varphi) + b \\ C(\tau_2) &= c(\tau_2) + b = -\frac{a}{2} \cdot cos(\varphi) + b \\ C(\tau_3) &= c(\tau_3) + b = \frac{a}{2} \cdot sin(\varphi) + b \end{aligned} \tag{6}$$

With this four selected points it is possible to calculate the correlation function and determine the phase φ and amplitude a of $s(t)$:

$$\varphi = atan\left[\frac{C(\tau_3) - C(\tau_1)}{C(\tau_0) - C(\tau_2)}\right] \tag{7}$$

$$a = \frac{\sqrt{[C(\tau_3) - C(\tau_1)]^2 + [C(\tau_0) - C(\tau_2)]^2}}{2} \tag{8}$$

The depth d is calculated by the following equation:

$$d = \frac{c \cdot \varphi}{2 \cdot 2\pi f} \tag{9}$$

where c is the speed of light and f is the IR modulation frequency [3].

B. Stereo depth camera model

Systems based on stereo vision use two sensors horizontally separated by known distance named baseline. They reproduce depth the way our binocular vision works by capturing the scene from different points of view. Solving the correspondence problem means giving a point in the image and finding the same point in another image. Figure 1 shows a model with two cameras, that have parallel optical axes. Observed point P is projected differently on the camera planes.

Depth information at the given pixel location is inversely proportional to the disparity defined as difference in horizontal coordinates of corresponding image points x_r and x_l. Using disparity d defined as $(x_r - x_l)$ is depth d calculated by following equation:

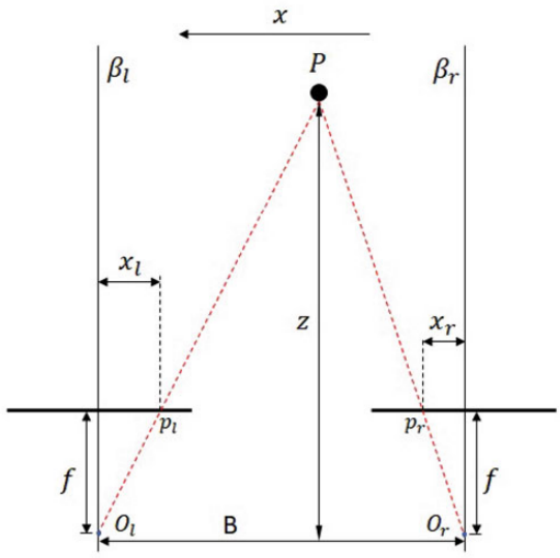


Fig. 2: Stereo system model [4].

$$d = \frac{B \cdot f}{(x_r - x_l)} \tag{10}$$

where f is focal length and B is the baseline [4].

Stereo cameras use computationally intensive algorithms for matching visual features to solving correspondence problem and measure depth. These systems will work well in most lighting conditions including outdoors. Specifically, depth camera ZED mini has two sensors separated by 12 cm and is based on this principle.

If there are fewer color and intensity variations in the image, stereo vision could be less effective. Active stereo vision relies on the addition of an optical projector that overlays the observed scene with a semi-random texture that facilitates finding correspondences, in particular in the case of texture-less surfaces like indoor dimly lit white walls. The current generation of RealSense D4xx cameras are able to pick up the slightest texture in a scene, especially in bright environments, and therefore works well outdoors.

In the case of scanning dynamic objects using the multi-camera system, the major benefit of this type of depth camera is that there are no limits to how many you can use in a particular space. It does not matter if other cameras point at the same scene with their projectors. To first order, all additional projectors actually improve the overall performance by adding more light and more texture [5].

C. Structured light camera model

Visual systems based on structured light need in addition to the sensor also some light source, that actively illuminates the scene with regular patterns. The projected pattern is distorted by the surface of the object. The pattern geometry is known, so the depth map of the scene can be easily estimated based on the distortion. The sensor recovers position and depth of the point

on the scene using equations 11 and 12 [4], where variables are shown in Figure 3. The shape can be reconstructed using the plane pattern and the triangulation principle.

$$z = \frac{B}{tan(\alpha) + tan(\beta)} \tag{11}$$

$$x = z \cdot tan(\alpha) \tag{12}$$

Projectors can effectively work in the infrared spectrum so all of this is invisible to the user. When using stripes as the projected pattern, the optical resolution of the depth map can be improved by the reduction of strips width. In extreme cases, it is limited by camera resolution and also by the wavelength of light.

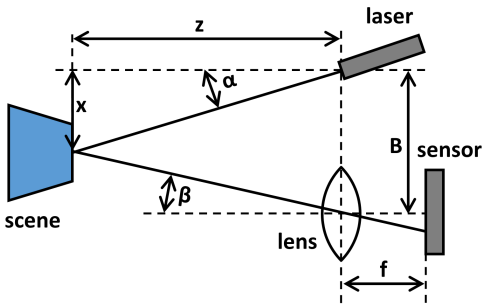


Fig. 3: Triangulation principle with single laser spot [4].

III. Related work

For our future work, we need to select technology, that allows capturing dynamic objects in the multi-camera system. As described in our previous study [6], by parallel scanning with ToF sensors is causing multi-camera interference. The same phenomenon occurs when structured light sensors are used because of overlaying projected patterns on the surface of objects. The technology that, in principle, does not suffer from interference in the multi-camera system is an active or passive stereo camera pair.

The accuracy of depth sensors of the same technologies was compared in several recent works [7]. The known methodologies for error estimation often needs a precise object with its ground truth model, which is difficult to get. The benefit of this work is a comprehensive comparison of all technologies in small distances only using testing patterns and surfaces. The several works [8], [9] show, that the ToF sensors are more accurate as of the stereo pairs. The practical contribution of this research is to evaluate the accuracy difference between these technologies and decide if the ToF and structured light sensors can be replaced with stereo pairs in multi-camera systems.

IV. Methodology

To compare we use two metrics for depth error estimation. Principally we are able to evaluate time variability of depth measurement against a flat surface for concrete distances. This deviation is represented as the noise of a depth sensor, which depends on the distance. In our case, we placed all cameras in distances 0.5 m, 0.7 m and 1 m from a flat surface. Comparison in the shorter or longer distance was not possible for the limitations of technologies. The data was collected for 10 seconds.

To estimate depth error we used a simplified technique based on the methodology described in the study [10]. A similar method is described and applied in the study [11] as a generalized method for depth error estimation for any device. To estimate accuracy we tried to compare point clouds generated from depth maps with ideal point clouds. Our points of interest are the corners of the chessboard of size 9×7 where each edge is 36 mm long. The same chessboard was modeled as the ideal reference point cloud. To find corners in RGB image the OpenCV method was used. Using equations of pinhole camera model for projection from image coordinate system to world coordinate system (X, Y, Z) we get the real point cloud captured by the camera. In this case, the Z-coordinate is depth at image pixel position (u, v). Intrinsic parameters of each camera are needed for calculation coordinates X and Y using following equations:

$$X = \frac{u - C_x}{f_x} Z \tag{13}$$

$$Y = \frac{v - C_y}{f_y} Z \tag{14}$$

The next step is rotation and translation estimation between the real and ideal cloud. The Coherent Point Drift algorithm is used as a global registration method and local precise fitting is realized using Iterative Closest Point. Euclidian distances in 3D space between real and ideal points represents Root Mean Square Error of measurement evaluated using the following equation:

$$RMSE = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (p_i - p_i')^2} \tag{15}$$

To avoid the error caused by not ideal construction of the three-dimensional pattern we used plane surface captured in three positions as shown in Figure 4.

V. Experiments and results

In our experiment, we compare four cameras of different technologies. KinectV2 as ToF sensor and RealSense SR300 as structured light sensor use infrared light and their cooperative usage is complicated. On the other hand, ZED MINI as passive and RealSense D415 as active stereo pair cameras are good candidates for using in multi-camera systems. The main parameters of all cameras are compared and summarized in Table I. All the parameters are available on product websites [12], [13] or comparison Table in [4].

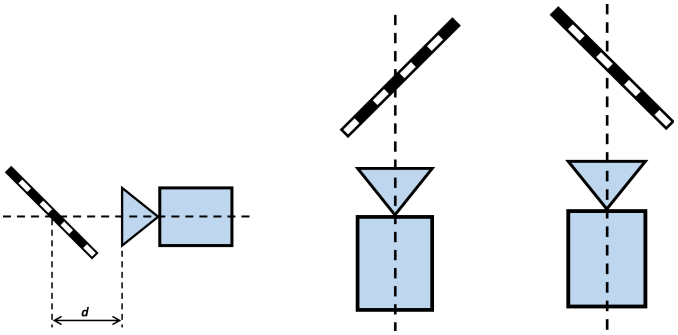


Fig. 4: Chessboard pattern positions: (**a**) Position A, side view. (**b**) Position B, top view. (**c**) Position C, top view.

TABLE I: Depth cameras parameters comparison [4], [12], [13].

Camera	**Depth sensor**				
	Technology	**DFOV**	**Max. Resolution**	**FR* [fps]**	**Range [m]**
RealSense D415	Active Stereo	72	1280x720	90	0.3 - 10
Kinect V2	ToF	70x60	512x424	30	0.5 - 4.5
ZED mini	Stereo	110	4416x1242	100	0.15 - 12
RealSense SR300	Coded Light	90	640x480	60	0.2 - 1.5

*Maximal FR value might depend on resolution used.
*DFOV - Diagonal Field of View.

A. *Noise measurement*

As shown in Table I, there are differences in possible resolutions of depth and RGB sensors. The resolutions of ZED MINI and D415 were set to Full HD except for depth map resolution of D415, which allows maximal 1280×720. SR300 allows using VGA and KinectV2 512×424 resolution of the depth map. Comparison of sensors noise for different distances is in Figures 5, 6, 7. The method of noise measurement is described in section IV.

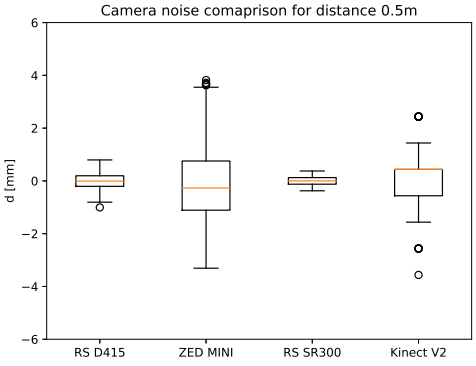


Fig. 5: Noise of depth sensors comparison for distance 0.5 m.

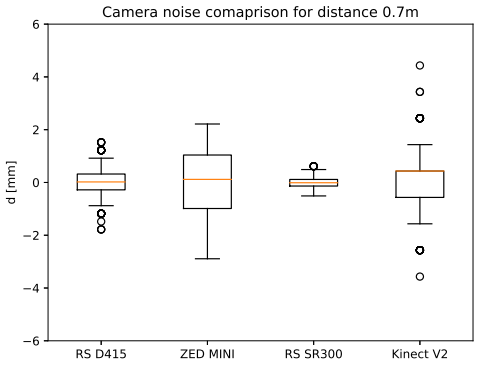


Fig. 6: Noise of depth sensors comparison for distance 0.7 m.

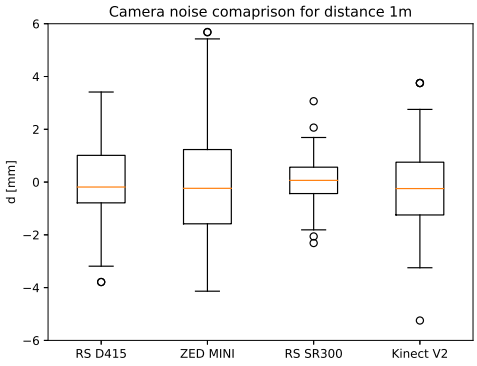


Fig. 7: Noise of depth sensors comparison for distance 1 m.

As expected, the noise increases with distance from the plane. We compare only the variable part of the signal to ignore the absolute distance offset of sensors. As seen in Table II, the best results were achieved using structured light technology. There is also a difference between active and passive stereo pair. The distance variation of ZED MINI as a passive stereo pair representative is higher than using an active stereo pair. This difference can be caused by problematic correspondence finding, what is improved by using an additional projector in the case of D415.

TABLE II: Standard deviation for multiple distances.

Camera	σ **for dist. [mm]**		
	500	**700**	**1000**
D415	0.307	0.639	1.303
ZED	1.499	1.343	2.180
KinectV2	1.151	1.267	1.375
SR300	0.124	0.253	0.716

B. *Ideal cloud fitting*

During point clouds comparing were used the same parameters of cameras as in the previous test. In the case of ZED MINI and RealSense D415, the testing pattern was captured from two distances of 0.5 m and 1 m. Error comparison from

different views as shown in Figure 4 is in the Table III. Registered point clouds captured by RealSense D415 is shown in Figure 8. Due to the same image resolution of D415 and ZED MINI, we can expect the same corner localization error, so the comparison of these two technologies can be considered as precise.

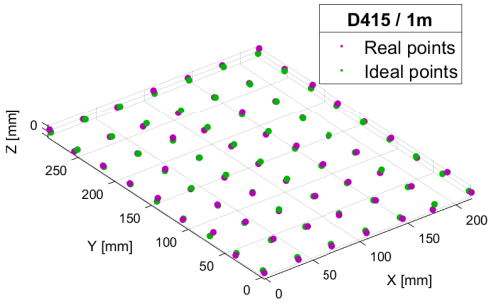


Fig. 8: Real point cloud captured by D415 and fitted to ideal cloud

TABLE III: RMS Error for multiple distances and positions using ideal cloud fitting.

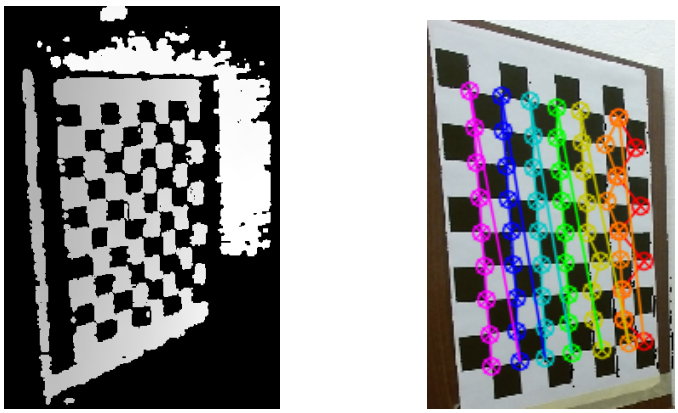
Camera	RMS Error for distance [mm]						
	Pos. A			Pos. B		Pos. C	
	500	700	1000	500	1000	500	1000
D415	1.323	1.881	3.273	1.535	3.137	1.288	3.106
ZED	1.439	2.881	4.046	2.450	5.342	1.519	4.358

During pattern capturing using SR300 and KinectV2, several complications occurred. Depth map captured by SR300 contains squared regions with unknown depth on black places of the chessboard as shown in Figure 9. That means we are not able to get the depth value of chessboard corners needed for point cloud reconstruction. The same effect occurred when using Kinect in distance more than 0.7 m from object. Distance deviation due to varying object reflection and associated ToF camera calibration are described in the study [14].

In the case of KinectV2, there is a limitation of image resolution, so the algorithm is not able to determine the right positions of corners on the chessboard, as shown in Figure 9. Due to these limitations, we are not able to compare all the technologies using ideal point cloud fitting.

C. Ideal plane fitting

In study [4] is the depth reconstruction accuracy estimated by fitting captured point cloud to ideal surface. Our approach is to capture the point cloud surface of the plane without chessboard in similar positions as in the previous case. The estimated error is then the difference between the ideal plane and real point cloud fitted to this plane, as shown in Figure 10. The comparison of all camera technologies is shown in Table IV.



(a) (b)

Fig. 9: Complications at getting real point clouds from sensors SR300 and KinectV2: (**a**) Unknown regions (black color) on depth map captured by SR300. The bright color indicates higher camera to object distances and darker color indicates the lower distances. (**b**) Corners localization error caused by lower camera resolution (KinectV2).

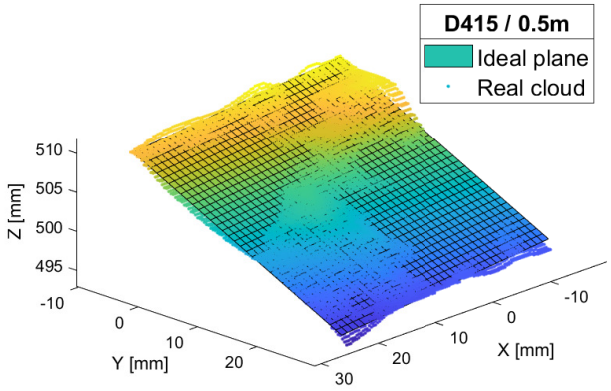


Fig. 10: Real point clouds captured by D415 and fitted to ideal plane.

TABLE IV: RMS Error for multiple distances and positions using plane fitting.

Camera	RMS Error for distance [mm]					
	Pos. A		Pos. B		Pos. C	
	500	700	500	700	500	700
D415	0.742	1.242	0.506	0.991	0.691	0.845
ZED MINI	0.516	0.8632	0.965	1.273	0.866	1.021
KinectV2	1.650	1.795	1.445	1.499	1.449	1.471
SR300	0.386	0.935	0.321	0.877	0.257	0.941

The resulting accuracy is not sensitive to corners localization error, so the results for corresponding positions in Tables III and IV are different.

VI. Conclusion

In our study, we tested depth cameras based on the known principles of sensing and tried to evaluate the depth measuring accuracy. The accuracy parameter is important for our future biomedical implementation. Due to limitations of structured light and ToF sensors we are not able to compare these technologies using ideal point cloud fitting. The comprehensive comparison was made by the ideal plane fitting. In future work, there is a possibility to use colored chessboard instead of black and white. Also, the precise 3D construction of square covered by chessboard pattern instead of a flat plane in different views is the potential way in complex comparison as described in [11].

According to the results of this work, the active stereo pair seems to be even more accurate than ToF sensor by usage in small distances. In terms of noise, the ideal camera to scanned object distance seems to be less than 1 m. Although, the structured light technology is most accurate but not suitable for parallel scanning mode as mentioned above. The difference in accuracy between active stereo pair and structured light is only 0.3 mm on average. Based on these findings we assume, that the accuracy of the active stereo pair is sufficient for obtaining 3D models of pediatric patients. For our future work, we prefer using the active stereo pair camera Intel RealSense D415. Our main goal was reached by selecting the most suitable technology for scanning in the parallel multi-camera system.

Acknowledgment

Results of this research are funded by APVV-15-0462: Research on sophisticated methods for analyzing the dynamic properties of respiratory epithelium's microscopic elements, APVV-17-0218 Investigation of biological tissues with electromagnetic field interaction and its application in the development of new procedures in the design of electrosurgical instruments.

References

[1] S. M. Islam, H. Mahmood, A. A. Al-Jumaily, and S. Claxton, "Deep learning of facial depth maps for obstructive sleep apnea prediction," Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018, pp. 154–157, 2019.

[2] David Bulczak, Martin Lambers, and Andreas Kolb, 'Quantified, Interactive Simulation of AMCW ToF Camera Including Multipath Effects', Sensors, vol. 18, no. 2, p. 13, Dec. 2017.

[3] R. Lange, 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. Department of Electrical Engineering and Computer Science at University of Siegen, 2000.

[4] S. Giancola, M. Valenti, and R. Sala, A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies. Cham: Springer International Publishing, 2018.

[5] A. Grunnet-Jepsen, J. N. Sweetser, P. Winer, A. Takagi, and J. Woodfill, 'Projectors for Intel® RealSense™ Depth Cameras D4xx', p. 14.

[6] J. Volak, D. Koniar, F. Jabloncik, L. Hargas, and S. Janisova, 'Interference artifacts suppression in systems with multiple depth cameras', in 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 2019, pp. 472–476.

[7] B. Langmann, K. Hartmann, and O. Loffeld, "Depth Camera Technology Comparison and Performance Evaluation," in ICPRAM, 2012, doi: 10.5220/0003778304380444.

[8] A. Vit and G. Shani, "Comparing RGB-D Sensors for Close Range Outdoor Agricultural Phenotyping," Sensors, vol. 18, no. 12, p. 4413, Dec. 2018, doi: 10.3390/s18124413.

[9] C.-Y. Chiu, M. Thelwell, T. Senior, J. Hart, and J. Wheat, "Comparison of Depth Cameras for 3D Reconstruction in Medicine," p. 15.

[10] L. E. Ortiz, V. E. Cabrera, and L. M. G. Goncalves, 'Depth Data Error Modeling of the ZED 3D Vision Sensor from Stereolabs', ELCVIA, vol. 17, no. 1, p. 1, Jun. 2018.

[11] L. Fernandez, V. Avila, and L. Goncalves, 'A Generic Approach for Error Estimation of Depth Data from (Stereo and RGB-D) 3D Sensors', MATHEMATICS & COMPUTER SCIENCE, preprint, May 2017.

[12] 'Stereolabs - ZED Stereo Cameras Web Site'. [Online]. Available: https://www.stereolabs.com/. [Accessed: 29-Oct-2019].

[13] 'Intel® RealSense™ Depth and Tracking Cameras', Intel® RealSense™ Depth and Tracking Cameras. [Online]. Available: https://www.intelrealsense.com/. [Accessed: 29-Oct-2019].

[14] M. Lindner and A. Kolb, 'Calibration of the intensity-related distance error of the PMD TOF-camera', presented at the Optics East 2007, Boston, MA, 2007, p. 67640W.