FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor’s thesis

Title:	Creation of 3D model using common depth sensors and its limitations
Student:	Filip Čacký
Supervisor:	Ing. Jakub Novák
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2022/2023

Instructions

Map the possibilities of creating a 3D model of a real object using various technologies based on depth sensors. Focus on ways to create a so-called 3D point cloud, scanning principles, their accuracy and their limitations (eg. different materials or sizes of objects). The result should be useful for choosing the technology in the state of feasibility study for industrial applications using depth sensors.

Goals:
1) Perform a search in the field of sensing methods using depth sensor technologies.
2) Assemble a measuring camera system.
3) Create a measurement methodology, including the specification of technology suitability.
4) Design algorithms for creating a 3D point cloud and working with it.
5) Test and evaluate the results of the proposed algorithms.
6) Visualize and discuss the results.

Electronically approved by Ing. Karel Klouda, Ph.D. on 1 February 2022 in Prague.

Bachelor's thesis

CREATION OF 3D MODEL USING COMMON DEPTH SENSORS AND ITS LIMITATIONS

Filip Čacký

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: Ing. Jakub Novák
May 12, 2022

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: Čacký Filip. *Creation of 3D model using common depth sensors and its limitations.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2022.

Contents

List of Figures

List of Tables

List of code listings

I want to thank my bachelor's thesis supervisor Ing. Jakub Novák, for his guidance, encouragement, and many meaningful discussions that helped me with the writing of this thesis.
I also want to thank my family for the unconditional love and support they have given me throughout my life and studies.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis. I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on May 12, 2022 . . . .

Abstrakt

Tato práce se zabývá zmapováním limitujících faktorů tvorby mračen bodů a meření vzdálenosti pomocí běžně dostupných hloubkových kamer.

Byl zkoumán vliv incidenčního úhlu, vzdálenosti měření, tvaru objektu, velikosti objektu a povrchových charakteristik na třech zařízeních, využívajících principy Time-of-Flight, pasivního stereo vidění a aktivního stereo vidění. Porovnaná zařízení jsou Basler ToF, OAK-D-Lite a Intel RealSense D415.

Testy odhalily silné a slabé stránky každého zařízení. Kamera Basler ToF má nejlepší strukturální rozlišení, ale je snadno ovlivnitelná povrchovou úpravou měřených objektů. Přesnost hloubky a strukturální rozlišení kamery Intel RealSense D415 jsou srovnatelné s Basler ToF, ale nejsou tak snadno ovlivněny povrchovou úpravou. OAK-D-Lite postrádá strukturální rozlišení potřebné pro přesnou rekonstrukci 3D objektů, ale zdá se být dobrou volbou pro použití ve vizuálně různorodých prostředí pro přibližný odhad hloubky scény.

Výsledky provedených experimentů jsou využitelné jako podklad pro studie proveditelnosti projektů strojového vidění s využitím hloubkových kamer.

Klíčová slova metrologická studie, hloubková kamera, porovnání hloubkových kamer, 3D rekonstrukce, Intel RealSense D415, OAK-D-Lite, Basler ToF, stereovize, time of flight

Abstract

This thesis is concerned with the mapping of limitations on creating point clouds and measuring depth using commonly available depth-sensing cameras.

The influence of an incidence angle, measurement distance, object shape, size and surface characteristics were investigated and measured on three devices, utilising Time-of-Flight, passive stereo vision, and active stereo vision. The compared devices are Basler ToF, OAK-D-Lite, and Intel RealSense D415.

The tests established the strengths and weaknesses of each device. The Basler ToF has the best structural resolution but is easily influenced by the surface characteristics of the measured objects. The depth accuracy and structural resolution of the Intel RealSense D415 are comparable to the Basler ToF but are not as easily influenced by a surface finish. The OAK-D-Lite lacks the structural resolution required for an accurate reconstruction of 3D objects. However, it seems to be a solid choice for an approximate scene depth estimation in visually complex environments.

The results of the carried out experiments are usable as a basis for feasibility studies on computer vision projects utilising depth-sensing cameras.

Keywords metrological comparison, depth camera, depth camera comparison, 3D reconstruction, Intel RealSense D415, OAK-D-Lite, Basler ToF, stereovision, time of flight

List of abbreviations

CMM	Coordinate Measuring Machine
ToF	Time of Flight
ICP	Iterative Closest Point
RANSAC	Random Sample Consensus
ROI	Region Of Interest
FOV	Field Of View
RGB	Red Green Blue
RGB-D	Red Green Blue Depth
BW	Black and White
SDK	Software Development Kit
NIR	Near Infra Red
LED	Light Emitting Diode

Introduction

Lately, the market is becoming increasingly saturated with readily available depth cameras of varying cost, accuracy, and intended use. Historically, short-range 3D reconstruction was prevalently done using mechanical or laser-based coordinate-measuring machines (CMMs) or laser-based hand-held scanners. While depth cameras are not intended to replace the devices mentioned above, their portability and accessible price enable a broad range of new technologies to be developed and widely adopted.

When imaging a scene using a monocular camera, information about the distance of scene objects from the camera sensor is lost due to projection to a 2D space. While it is possible to reconstruct the depth of a scene from acquired images, it often requires a large number of images, heavy preprocessing, and a long run time. Thus it is not suited for applications when real-time depth sensing is required or computing power is limited.

Depth sensing systems are often deployed in many fields for various use cases. Some such use cases include, but are not limited to, quality control of parts created by additive manufacturing, inspections of buildings, reconstruction of historical artifacts, terrain mapping, navigation, and localization of robotic agents in their working environment.

In contrast to CMMs and hand-held scanners, the ability of depth cameras to reproduce accurate 3D models is not advertised by their manufacturers. The ability to capture small details is not advertised at all, while the depth estimation accuracy is often reported as a deviation of mean distance in a central ROI to a perpendicular plane from the ground truth. This metric, however, does not represent the device's ability to capture complex scenes accurately. Because of this, it is hard to judge the suitability of different devices to carry out required tasks.

This thesis is concerned with conducting a metrological comparison of selected depth cameras, utilizing different principles for estimating the depth of the scene. The comparison will be conducted on both single-view and multi-view scenes. This metrological comparison will then serve as a basis for feasibility studies for computer vision projects utilizing depth cameras. The methodology for metrological comparison is in part based on the ISO-10360:13 standard.

Chapter 1
Theoretical background

Computer vision is concerned with the extraction of information from visual inputs.

1.1 Image segmentation

Image segmentation is the process of separating different parts of an image in order to extract an ROI in order to simplify the solved task. The segmented parts are then further used for various measurements, feature extraction, or more complex tasks such as object tracking.

A simple example of an image segmentation algorithm is thresholding, which extracts parts of an image whose pixel values lie within a certain predefined threshold.

1.2 Image filtering

Image filters are used to change an image on a pixel level. Different filters can accomplish a broad range of results. For example, highlight the present edges or corners or suppress the noise in an image by blurring.

An example of an edge detecting filter is the Sobel filter. A mean filter can be used to blur an image in order to reduce noise.

1.3 Feature detection and description

Feature detection and description methods are used to extract and mathematically describe interesting regions in an image. Such interesting regions may be parts of an image where, for example, a sudden change of brightness or colour occurs. Those regions are referred to as key points.

The description process is used in order to discern or match different key points from each other in different images. The problem of finding corresponding points in multiple images is referred to as the correspondence problem and is the fundamental problem of computer vision.

SIFT, SURF and ORB are examples of algorithms used for feature detection and extraction.

1.4 Intrinsic camera parameters

Intrinsic parameters of a camera are focal length, principal point offset, and shear coefficient. The intrinsic camera matrix transforms 3D camera coordinates into 2D homogeneous image coordinates. The ideal pinhole camera models this projection.

The intrinsic matrix I is parameterized as

$$I_{3\times3} = \begin{pmatrix} f_x & s & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.1}$$

Where f_x and f_y describe the focal length of the camera in counts of pixels. Two lengths are used due to imperfect (in other words, not square) dimensions of the sensor pixels. x_0 and y_0 describe the offset of the principal point from sensor origin. s is the shear coefficient [1].

1.5 Epipolar geometry and constraint

Epipolar geometry is the basis of stereo vision. It establishes the relations between two images of a single scene captured from different perspectives. The epipolar constraint simplifies the problem of finding corresponding points in those images (known as the correspondence problem).

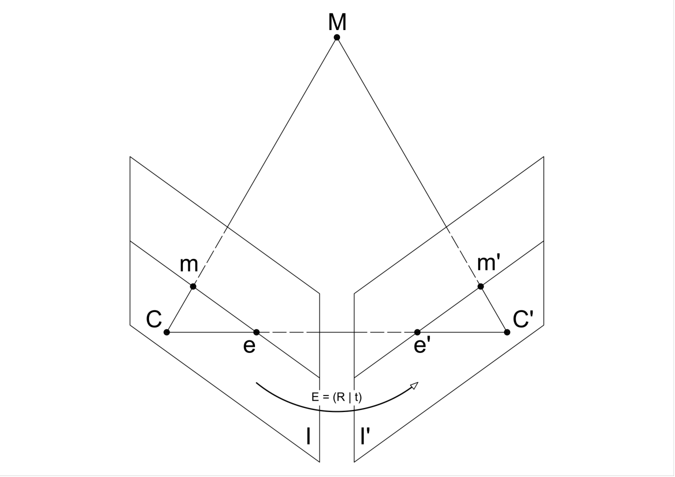


Figure 1.1 Scene shot from two perspectives described by epipolar geometry.

Figure 1.1 represents two views of the same scene. C and C' represent optical centres of the first and second cameras, I and I' their respective image planes, where rigid transformation $(R \mid \vec{t}) := E$ describes the rotation and translation of the second camera relative to the first and is referred to as the essential matrix of the stereo system. If measurement in pixels is required, a fundamental matrix F is used instead, which, in addition to the rigid transformation between the two camera sensors, also contains the intrinsic parameters. The distance between optical centres is referred to as the baseline. M represents a point in the scene. m and m' are projections of the point M on image plane I and I' respectively. Projection of camera centres C to image plane I' gives us a point called the epipole e of image plane I, analogically for e'.

The line is defined by a projection of point M on an image plane and an epipole, is called the epipolar line corresponding to the projection of point M.

Corresponding epipolar lines, created by the same point M on each of the image planes, are referred to as a conjugate epipolar pair.

The epipolar constraint establishes that a projected point in one image plane must lie on the corresponding epipolar line in the second image plane. Therefore it simplifies the solution of the correspondence problem. This second epipolar line can be easily determined by transforming the first one by the fundamental matrix F. [2, 3]

1.6 Stereo rectification

Image rectification is a projective transformation of a pair of stereo images to a common image plane. This is particularly useful, as solving the correspondence problem then becomes much more manageable.

Rectifying a stereo image pair also transforms the conjugate epipolar line pairs such that each pair becomes colinear and parallel to one of the image axes (usually the horizontal axis).

The advantage of rectification is that the search of corresponding points is reduced from 2D space to search along the horizontal axis of the transformed images. [4]

1.7 Disparity

Two rectified images used for calculation of disparity are shown in figure 1.2.

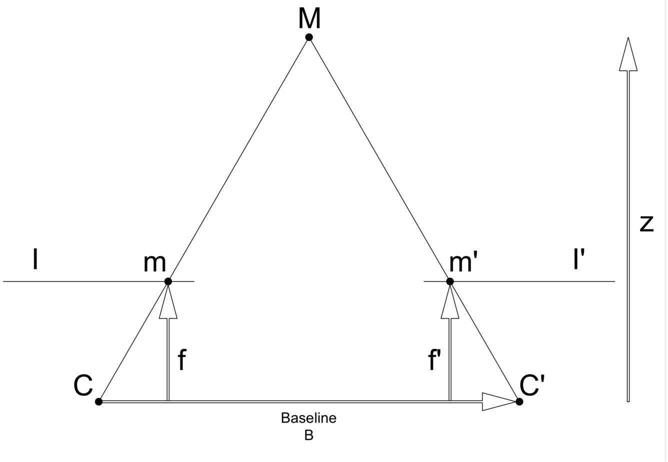


Figure 1.2 A diagram of two rectified images

After determining the corresponding key points in both images, the disparity can be calculated. The stereo disparity is defined as the offset between a point m in image I and point m' in image I' calculated along the horizontal axis. The calculated disparity will be positive in the left image and negative in the right image.

Having obtained this difference, the depth Z in the specified pixel position can be obtained by the following equation.

$$z = \frac{B \times f}{m_x - m'_x} \tag{1.2}$$

Where B is the baseline in cm and f is the focal length in counts of pixels [5].

In practice, key points are matched by blocks of predetermined size. Larger blocks produce fuller maps with a lower amount of noise, while smaller blocks produce more accurate depth maps.

For best results, certain soft constraints should be complied with. Such as low disparity gradient and uniqueness of key points.

There are, of course, many other difficulties in producing correct depth maps by stereo matching, such as occlusions in one of the views or uncertain correspondences.

1.8 Depth image

The depth image is an $n \times m$ matrix, where each pixel within the image plane contains the distance to a 3D point in the scene projected to that position. Invalid depth is often represented by a zero.

1.9 Confidence image

A confidence image is an image of the same size as the depth image, where each pixel contains the confidence the device has in estimating the depth value at the same coordinates. Higher values and lower values correspond to higher and lower confidence, respectively. The generation of a confidence image is device-specific.

The primary usage of a confidence image is to remove values from a depth frame based on the likelihood that they are incorrect. This can be useful during scanning for the removal of incorrectly measured points or especially useful in robotic applications, where real-time depth sensing is often required and where incorrect or ”hallucinated” depth may have severe effects on the operation of the device.

1.10 Temporal filter

A temporal filter increases depth map consistency and accuracy by manipulating pixel values based on previous depth frames. Either by imputing previous values into invalid pixels or pixels with low confidence, or by averaging the pixel difference of two consecutive images, thus reducing the amount of noise in static scenes.

It is best suited for scenes with low movement, as it introduces blurring and or smearing effects on the depth map [6].

1.11 Hole filling filter

A hole filling filter is used to impute values to invalid parts of the depth map by imputing from adjacent pixels. There are different approaches to the selection of value that will be imputed from the neighbouring pixels, such as direction-based imputing, imputing the furthest value by depth, or the closest one [6].

1.12 Spatial filter

A spatial filter is used to smoothen the captured scene by doing a series of horizontal and vertical passes, preserving edges. Depending on the implementation, it may also perform hole filling during the passes on the depth map by imputing values from adjacent pixels [6].

1.13 Point cloud

A point cloud defines a 3D scene as a set of points, where each point has an x, y, and z coordinate. In addition, a colour of a point, surface normal vectors, or detected features may also be specified.

Point clouds may be organized or unorganized. Organized point clouds resemble a matrix. Such point clouds have to contain invalid points (in positions where depth could not be determined), and are stored in either a row-major or column-major ordering. In contrast, unorganized point clouds contain only an unordered set of valid points.

Depth image can be extracted from an organized point cloud by extracting only the z component of the ordered coordinates. This extraction is impossible with unorganized point clouds due to missing points.

An example of a file format in which point clouds can be stored is the PCD file format. [7]

1.14 Scene reconstruction from depth image

If intrinsic parameters, introduced in section 1.4, of a depth image are known, The 3D scene can be reconstructed from the depth data.

Given depth value d at (u, v) image coordinate, the corresponding 3D point is:

$$z = d \tag{1.3}$$

$$x = \frac{(u - c_x) \times d}{f_x} \tag{1.4}$$

$$y = \frac{(v - c_y) \times d}{f_y} \tag{1.5}$$

Such reconstructed scene will be in the coordinate system of the camera used to produce the depth image, i.e. the origin vector $(0, 0, 0)$ will be placed in the centre of the camera sensor. [8]

1.15 K-dimensional tree

K-d tree is a multidimensional space-partitioning data structure. It was first presented in 1975 by Jon Louis Bentley in [9]. It can be described as a multidimensional binary search tree and is used for efficient nearest neighbour searches, among other types of queries.

3-dimensional k-d trees are often used for nearest neighbour queries in point clouds and are implemented, for example, in the Open3D library [1] or in scikit-learn [2].

1.16 Iterative Closest Point (ICP)

Iterative Closest Point is most often used for the alignment of polygonal 3D models or point clouds but can also be applied to geometric data such as polylines, parametric and implicit curves, and surfaces. Such a process is also referred to as registration. In the 3D vision, it is used extensively to align point clouds taken from different perspectives to a common space.

ICP is an algorithm for finding such rigid transformation $T = (R \mid \vec{(t)})$ between two point clouds, where R is a matrix determining the rotation and $\vec{t}$ is the translation vector, that when each point from the first point cloud is transformed by T, the distance between the two clouds is minimal.

Given two point clouds P_1 and P_2 and a set of correspondences between these clouds $C = \{(p_a, p_b) \mid p_a \in P_1 \wedge p_b \in P_2\}$, the algorithm can be simply described as the following optimization problem.

$$\text{minimize}\,\text{ICP}(T) = \frac{1}{\#C} \sum_{(p_a, p_b) \in C} ||T \cdot p_a - p_b|| \tag{1.6}$$

The ICP algorithm always monotonically converges to the nearest local minimum of a mean-square distance metric. Where the convergence is rapid during the first iterations [10].

[1] http://www.open3d.org/docs/latest/tutorial/Basic/kdtree.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KDTree.html

Due to that, the objects should already be closely aligned, and the set of correspondences be well constructed. Otherwise, a possibly incorrect transformation will be computed, as is shown in figure 1.3.

Such issue can be alleviated by, for instance, a predetermined camera trajectory, for which every sampling position is known, initializing the algorithm with a transformation created by use of odometry, or by a transformation computed from extracted key points for which correspondence is well known.

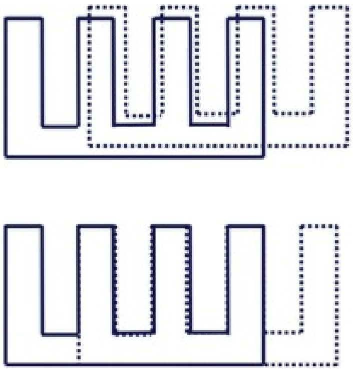


Figure 1.3 Uncertain ICP correspondence may result in incorrect registration [11].

1.17 Visual odometry

Odometry, in the sense it is being used in this thesis, is a term originating from robotics. It refers to the process of estimating the trajectory of a robotic agent by the use of sensors.

Similarly, visual odometry is the process of estimating a trajectory from a series of monocular or stereo images. Monocular odometry has the disadvantage of being unable to determine the scale of a translation between two positions. This disadvantage can be alleviated by supplying scale from an external source such as an odometer or scale calculated using two consecutive depth images.

1.18 RANdom SAmple Consensus (RANSAC)

RANSAC, first introduced by Fischler and Bolles in [12], is an algorithm designed for fitting a model (for example, a plane) to data containing a large number of errors. Due to that, it can also be used as an outlier detection algorithm. It can be naively summarised by the following pseudo-code listing 1.1.

Code listing 1.1 Pseudocode of RANSAC

```
def RANSAC(point_cloud, model, threshold, tolerance, max_it):

        inliers = []
        i = 0

        do:
                subset = sample(point_cloud, model.minimum_points)
                model.fit(subset)
                distances = model.calc_point_distance(point_cloud)
                inliers = point_cloud[distances < tolerance]
                i++

        while (inliers.size() / point_cloud.size()) > threshold &&
                  i < max_it

        return model.fit(inliers)
```

Chapter 2
Research

A depth camera is a camera capable of measuring the distances of objects in the shooting scene from the camera sensor (referred to as depth of the scene, depth image, depth map, or range image).

The process of shooting a scene with a depth camera is often referred to as 3D surface measurement, depth sensing, or depth mapping. [13]

2.1 Depth camera use cases

A wide range of industrial applications and academic research areas concerning depth cameras can be found. Such applications include, for instance, quality inspection [14, 15, 16, 17], biometrics [18], human-machine interaction [19] and robotics [20, 21].

[14] is dealing with visual inspection of reinforced concrete (RC) beam surface using an Intel RealSense D435 camera [1] . Two groups of RC beams were built, control beams with no external defects and beams with various surface deficiencies. Using the camera to produce depth maps of the beam's surface, they were able to qualify surface damage (roughness and cracks) correctly by comparing acquired depth maps with pre-established ground truth.

Similarly [15] used a structured light sensor mounted on a robotic vehicle to aid in the localization of external defects on an aeroplane hull. They were able to extract positions and dimensions of external hull defects by segmenting irregularities in produced point clouds.

[21] is concerned with the guidance of a robotic welding arm. Using point clouds generated by low-cost depth cameras, such as Intel RealSense D435 or ZED Mini [2], they created a point cloud processing pipeline for estimating welding seam positions with satisfying precision and flexibility.

2.2 Limitations of depth sensing

Several influencing factors limit the usability of depth cameras. The majority of them are imposed by the scanned material's surface, mutual position of the camera, and measured object or operating environment. Those limiting factors may include the following, depending on the technology used.

- high incidence angle, either of the measured object or situated somewhere in the scene itself, thus possibly causing multiple reflections to occur

[1] https://www.intelrealsense.com/depth-camera-d435/

[2] https://www.stereolabs.com/zed-mini/

- poor lighting
- aggressive lighting in the operating optical wavelength band of the used sensor
- high reflectivity of surface finish, either of the measured object or in the scene
- high dispersivity of the measured surface

2.3 Stereo cameras

Stereo depth cameras consist of two matrix sensors placed at known positions so that the fundamental matrix F of this system is known.

Estimation of scene depth by a stereo camera can be described by the following steps. First, an image of the scene is taken by each camera. The stereo image pair is then rectified. In the next step, key points are detected and matched. Having a list of key points from both images and their correspondences, disparity and depth can then be calculated as described in section 1.7.

The main limitation of stereo depth vision is that the scene for which depth is to be estimated must be visible by both sensors. Therefore there is a vertical band on both sensors, where no depth can be perceived. This limitation is illustrated in figure 2.1. D_{min} denotes the minimum distance at which depth can be perceived, as it is the closest point to the camera sensors, where projected pixels start to overlap. B_1 and B_2 describe the two areas where no depth can be perceived. [22]

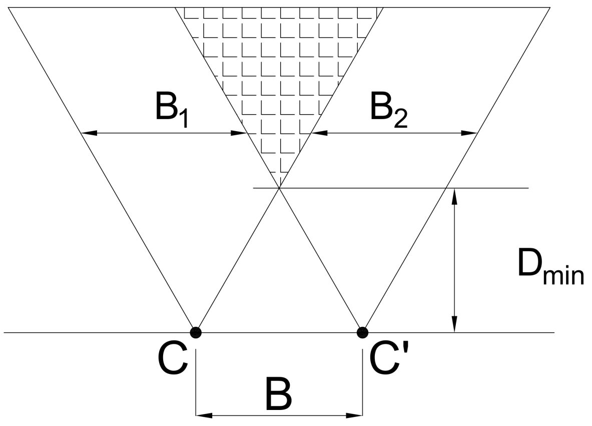


Figure 2.1 Limitation of stereo vision

2.4 Passive and active stereo cameras

In order to extract key points from images, feature detection and key point extraction methods are used, described in section 1.3. Passive cameras rely only on features present in the stereo images. Therefore their accuracy strongly depends on the complexity of the scene itself. Poorly lit or textured scenes will not contain many distinctive features. Therefore, the disparity matching algorithm will not have enough correspondences to calculate depth via dense matching. In that case, active stereo cameras need to be used.

Active stereo cameras combine the two sensors with an IR dot projector that illuminates the scene. The projected dots themselves are then used as features in the images. [23]

Multiple active stereo cameras shooting the same scene will not negatively affect each other, in contrast to technologies described further. The generated depth maps may be denser and more accurate due to a denser array of dots (features) being projected on the scene,

A comparison of produced depth maps by an active and a passive camera can be seen in figure 2.2. Black spots on the disparity maps on the right side of the figure are places where no corresponding features in stereo image pair could be found.

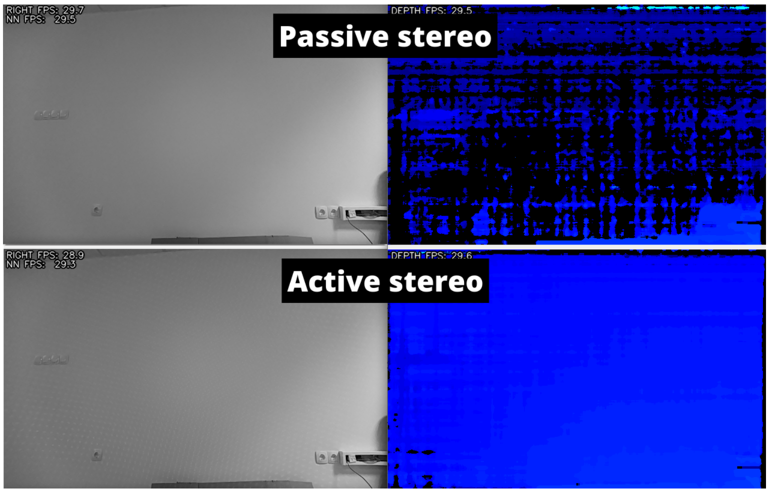


Figure 2.2 Visual comparison of passive and active stereo camera depth maps [23]

2.5 Structured light

Depth cameras utilizing structured light to determine scene depth consist of an active structured light projector and a camera.

The light projector emits a specific 2D pattern, or a sequence of patterns, predetermined by the nature of the scene. Scenes with movement may often require a single shot of the pattern, while a sequence of patterns and images may be used for a static scene and produce a more accurate depth map. The camera then captures the illuminated scene.

The captured 2D pattern will be nearly identical to the projected one if no objects are present in the scene. If an object is present, then the projected pattern will appear deformed, as can be seen in figure 2.3.

The depth map is then determined from the distortion of the projected pattern, as seen by the camera. For this, several different techniques are used, each depending on the emitted pattern. The most straightforward of which is triangulation. Triangulation is used, for example, with a full-frame rainbow pattern, which consists of spatially varying wavelength illumination, similar to the one depicted in figure 2.3. For this principle to work, the mutual position of the projector and camera needs to be known, in addition to the viewing angle of the scene. [13]

Structured light depth sensing is not well suited for applications, where multiple cameras utilizing the same pattern or sharing an optical wavelength band are required to operate in close vicinity of each other, thus possibly mangling each other's patterns if no synchronization is carried out.

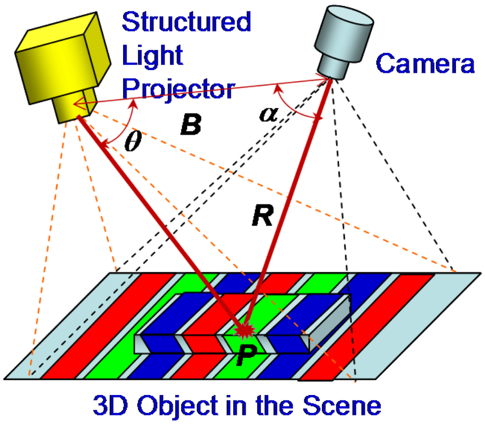


Figure 2.3 Camera system using structured light [13].

2.6 Time of Flight (ToF)

Time of Flight camera estimates the scene depth by measuring the time in which a modulated light is emitted, reflected, and registered by a sensor. It consists of a light source that illuminates the shot scene with a modulated light source (most often in the near infra-red spectrum) and an infra-red camera sensor.

Due to the way a ToF camera operates, it is not well suited to imaging scenes in which objects are positioned at high incidence angles. That will often lead to bad depth estimation or artefacts in the depth map. Other badly influencing factors of ToF measurement include, for example, low reflectivity of measured objects, aggressive ambient light or high infra-red noise, and other ToF cameras operating in the same scene, the last of which can be alleviated by synchronization. [24]

2.7 Compared cameras

The following cameras were compared in the metrological study, their parameters are displayed in table 2.1, as advertised by their manufacturers. The displayed values may depend on the specific settings the devices are initialized with.

2.7.1 Intel RealSense D415

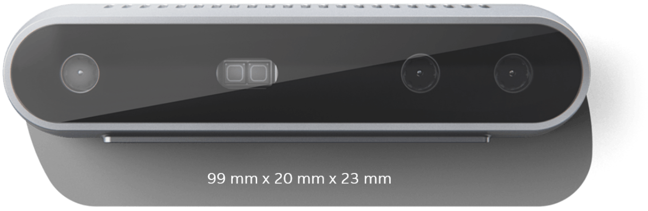


Figure 2.4 Intel RealSense D415 [25]

Intel RealSense D415, first launched in 2018, is an active RGB-D stereo camera. It consists of an IR dot projector, IR sensitive stereo pair, and a 1080P RGB sensor.

Intel also develops an open-source software development kit, the Intel RealSense SDK 2.0 [3], used for setup and acquisition of data. The SDK also offers a range of post-processing functions for the captured data.

The camera supports a range of different settings and presets [4]. The High Accuracy preset generates lower density depth maps with higher confidence, and the High Density preset generates higher density maps, at the expense of an increased amount of noise present in the depth maps.

2.7.2 Basler ToF

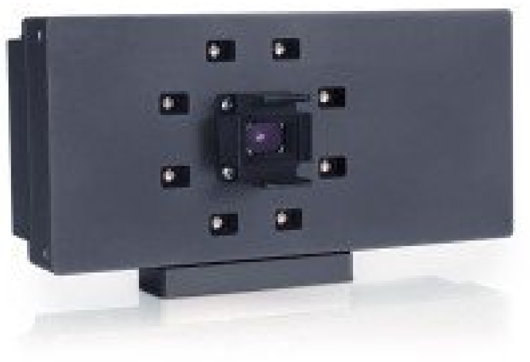


Figure 2.5 Basler ToF [26]

Basler ToF, first released in 2016, is, according to Basler, the first industrial ToF Camera with 480P resolution in the mid-range price segment. It has since been superseded by Basler Blaze 101 [5], with nearly all of the mentions of its existence removed from the official Basler websites. It has a standard GigE interface and a fully GenICam compliant interface, so it can be easily integrated into any computer vision project. It consists of 8 high-power NIR LEDs and a NIR camera sensor. It operates on the principle of pulsed ToF.

2.7.3 OAK-D-Lite

OAK-D-Lite is the newest addition to a series of depth cameras developed by Luxonis together with the OpenCV team. Its development has been backed on Kickstarter [6]. The camera consists of a BW stereo pair and an RGB camera sensor.

[3] https://github.com/IntelRealSense/librealsense

[4] https://github.com/IntelRealSense/librealsense/wiki/D400-Series-Visual-Presets

[5] https://www.baslerweb.com/en/products/cameras/3d-cameras/basler-blaze/

[6] https://www.kickstarter.com/projects/opencv/opencv-ai-kit-oak-depth-camera-4k-cv-edge-object-detection

Figure 2.6 OAK-D-Lite [27]

The interesting thing about this device, in contrast to the others, is that it is built around an Intel Movidius Myriad X VPU, allowing it to run complex workloads, such as neural networks, on-device without needing a performant host.

Luxonis develops an open-source software development kit DepthAI [7], used for setup and calibration of the device, in addition to the acquisition of frames from the device and deployment of custom payloads on the camera.

Currently, there are two different presets for depth estimation - High accuracy and High density.

Table 2.1 Camera specifications as advertised by the manufacturers [25, 26, 27].

	RealSense D415	Basler ToF	OAK-D-Lite (AF)
Technology	Active stereo	Time of Flight	Passive stereo
Resolution	720P	480P	480P
Max frame rate	90 FPS	20 FPS	200 FPS
Depth Accuracy	under 2% at 2 m	1 cm (distance unspecified)	unspecified
Field of View (H x V)	65° × 40°	57° × 43°	73° × 58°
Ideal Range	.5 m to 3 m	0 m to 5 m (or 13 320 mm)	.20 m to 19.1 m

[7] https://github.com/luxonis/depthai

Chapter 3

Measurement methodology

Several distinct experiments were carried out. Namely the flat-form distortion test, sphere shape test, sphere spacing test, distance estimation test, concave surface profile estimation test, and multi-view fusion test. The first three of which were inspired by, but have not conformed to precisely, the ISO 10360-13 standard on acceptance and reverification tests for optical 3D coordinate measuring systems. The reason for which is stated in section 3.1.2.

The general methodology each of these experiments followed can be described by figure 3.1.

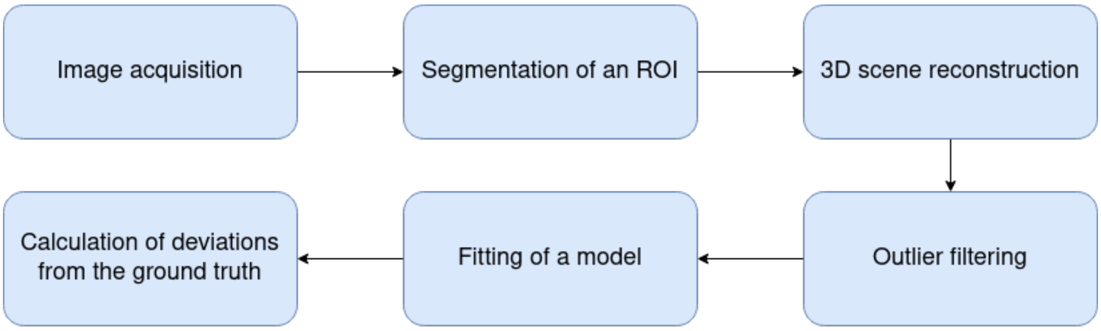


Figure 3.1 General measurement methodology

3.1 ISO 10360-13

The ISO standard 10360-13 [28], published in 2021, describes a number of acceptance and reverification tests for optical 3D measurement systems. An example of another similar standard would be VDE/VDE 2634 BLATT 2 [1]. ISO 10360-13 is, however, the only internationally accepted standard for the measurement of the accuracy of optical 3D CMMs, i.e., depth cameras.

The purpose of an acceptance test is the verification of the manufacturer's claimed performance of their devices, while reverification tests are carried out in order to ensure the long-term accuracy and usability of said devices in their working environments.

It describes the following tests in addition to the minimal and maximal sizes and positions of the artefacts.

- Probing characteristics measured on a sphere - estimation of a sphere shape

[1] www.vdi.de/en/home/vdi-standards/details/vdivde-2634-blatt-2-optical-3-d-measuring-systems-optical-systems-based-on-area-scanning

- Distortion characteristics measured on a pair of spheres situated in the measurement volume - estimation of sphere spacing
- Flat-form distortion error measured on a plane

3.1.1 General procedure of the ISO 10360-13 tests

For each of the tests, the total sensor measurement volume (or the required measurement volume) is divided into eight similarly sized regions. The measured artefact is then positioned in a different region (or their combination) for each of the carried out measurements in the tests. The error is then calculated across all the measurements.

3.1.2 Deviations of carried out tests from the standard

The main goal of the carried out tests was to compare the available devices on their ability to capture the used artefact accurately and to determine how the accuracy changes with respect to measurement distance, incidence angle, different materials of artefacts or a combination of thereof.

By the methodology of ISO 10360-13, the characteristics of the tested device should be measured in the entirety of its measuring volume to assure uniform accuracy in the entire designated working area. As this was not the goal in mind, the methodology could thus be simplified, and ultimately different attributes were observed. They should, however, be well representative of the results each device would be capable of achieving in the entire ISO test in comparison with each other.

Each device was compared under the same conditions, distances, and incidence angles. The observed attributes were, namely, the accuracy with which a device can capture spherical objects, the accuracy with which a device can capture planar surfaces and the accuracy with which a device can measure concave areas on an object.

3.2 Measured artefacts

Several different artefacts had to be obtained and manufactured to capture a dataset on which deviations would be measured.

3.2.1 Spherical artefacts

Five spheres made of wrought iron with a diameter of 30 mm characterized by a slightly shiny surface, five spheres made of wood with a diameter of 30 mm characterized by a rough matte surface and one sphere made of aluminium with 40 mm diameter and a gleaming surface was used.

The spheres were placed in a custom made holder, shown in figure 3.5, figure 3.3 and figure 3.4, at heights of 80, 60, 100, 90 and 70 mm respectively. Table 3.1 shows the respective centre-centre distances of the spheres when placed on the holder.

3.2.2 Planar artefacts

An 800 mm by 600 mm wooden board characterized by a rough matte surface hung from a supporting structure, and a smooth exposed concrete wall with many distinct key points was used for tests with planar surfaces.

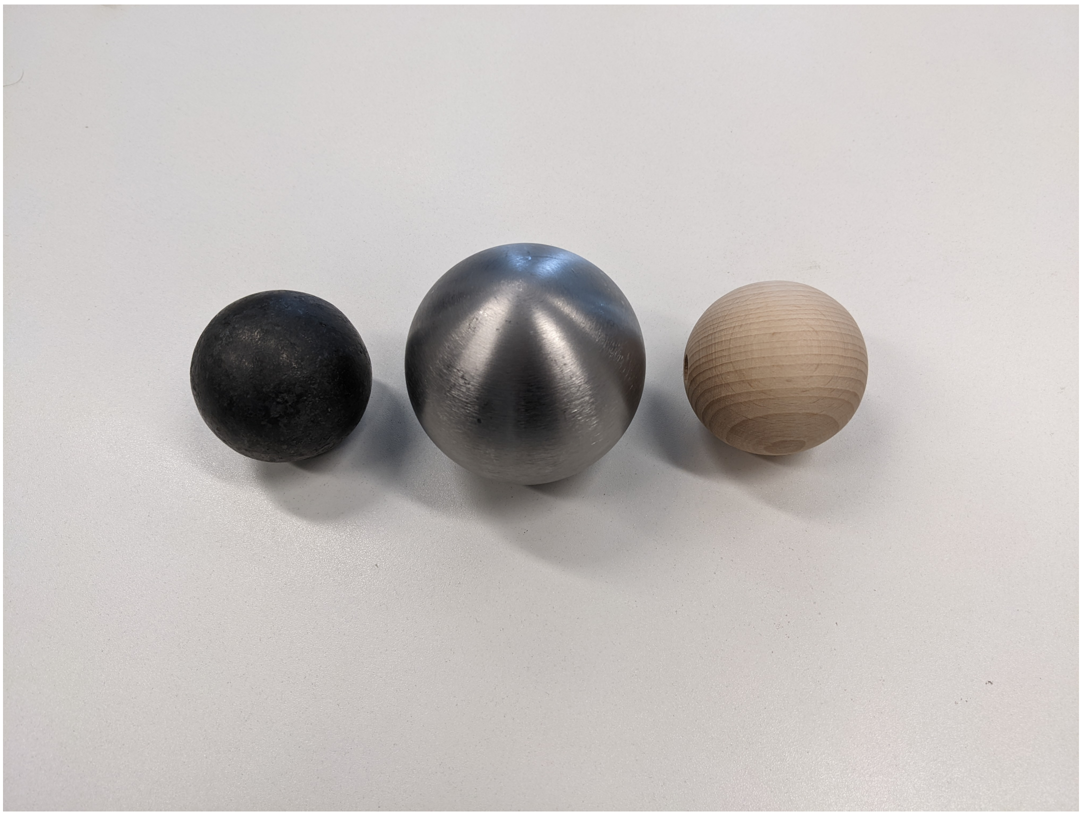


Figure 3.2 Wrought iron sphere, aluminium sphere, and wooden sphere with a radius of 30 mm, 40 mm, and 30 mm, respectively

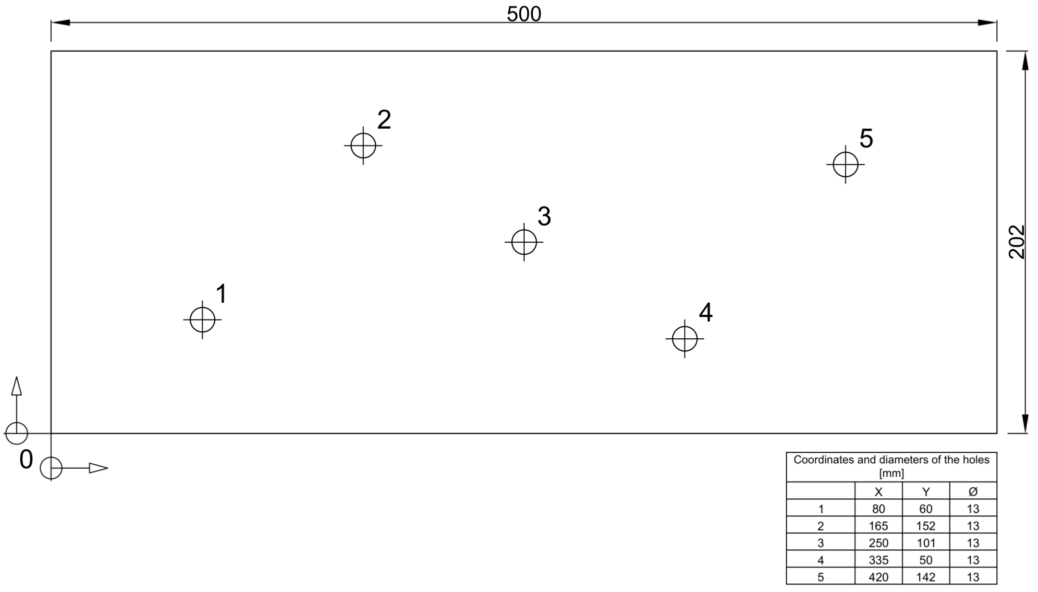
3.2.3 Cuboidal artefacts

Six cuboidal objects, shown in figure 3.8, with a cutout in the middle of one of the faces, were 3D printed. The artefacts were 60 mm by 60 mm wide, 30 mm high with cutouts of 10x10, 10x20, 10x30, 20x10, 20x20, 20x30 mm. The artefacts were 3D printed from black, opaque PLA filament.

Passive stereo vision is not expected to be able to reconstruct these artefacts accurately because the 3D printed PLA lacks distinct key points and is entirely uniform.

Table 3.1 Centre-centre distance of spheres when placed on the sphere holder artefact [mm], numbered from left to right.

Sphere number	S1	S2	S3	S4	S5
S1		127	176	255	350
S2	127		107	200	255
S3	176	107		100	178
S4	255	200	100		127
S5	350	255	178	127	



Coordinates and diameters of the holes [mm]			
	X	Y	Ø
1	80	60	13
2	165	152	13
3	250	101	13
4	335	50	13
5	420	142	13

Figure 3.3 Sphere holder artefact drawing from the top

3.2.4 Free-form artefacts

A guitar sound effect pedal was used as a free-form artefact. The pedal is made of cast metal with a shiny metal button and switch and a shiny plastic potentiometer, A corresponding 3D triangle mesh model was then reconstructed in OpenSCAD in order to obtain a reference. The artefact is shown in figure 3.9 and the 3D model in figure 3.10.

3.3 Used mathematical notation

The following notation is used in the definitions of the errors.

- $\#P$ refers to the size of set P.
- P_i is the i-th item in the set P.
- p_x and p_y refer to the x and y component of the point p, respectively.
- $||p||$ is the L2 norm of the point p.
- $|r|$ is the absolute value of r.
- (x, y, z) is a point in the 3D space.

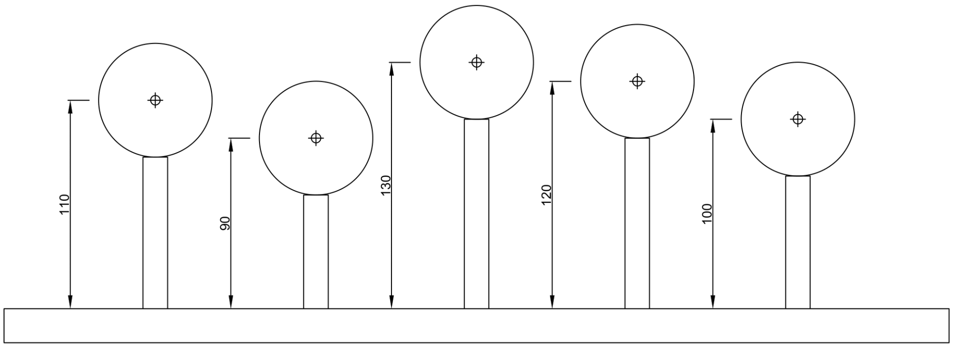


Figure 3.4 Sphere holder artefact drawing from the side

3.4 Tests on spherical artefacts

This test suite measures the ability of each of the devices to estimate the shape of spherical objects. The main interest was testing how a rapidly increasing angle of incidence affects the depth accuracy of a sensor.

Given a point cloud of a scene containing spherical artefacts, points belonging to each of the artefacts are segmented into separate point clouds.

A model of a sphere defined by the central point $c = (x, y, z)$ and radius r is then fit to each of the point clouds using RANSAC.

3.4.1 Sphere shape error E_r

The measured error is defined as follows.

Let R_T be a set of actual radii of the used spheres and R be a set of radii estimated by RANSAC.

$$E_r = \frac{1}{\#R} \sum_{i=0}^{\#R} |R_{Ti} - R_i| \tag{3.1}$$

In other words, E_r is the mean absolute error of estimated radii from actual radii.

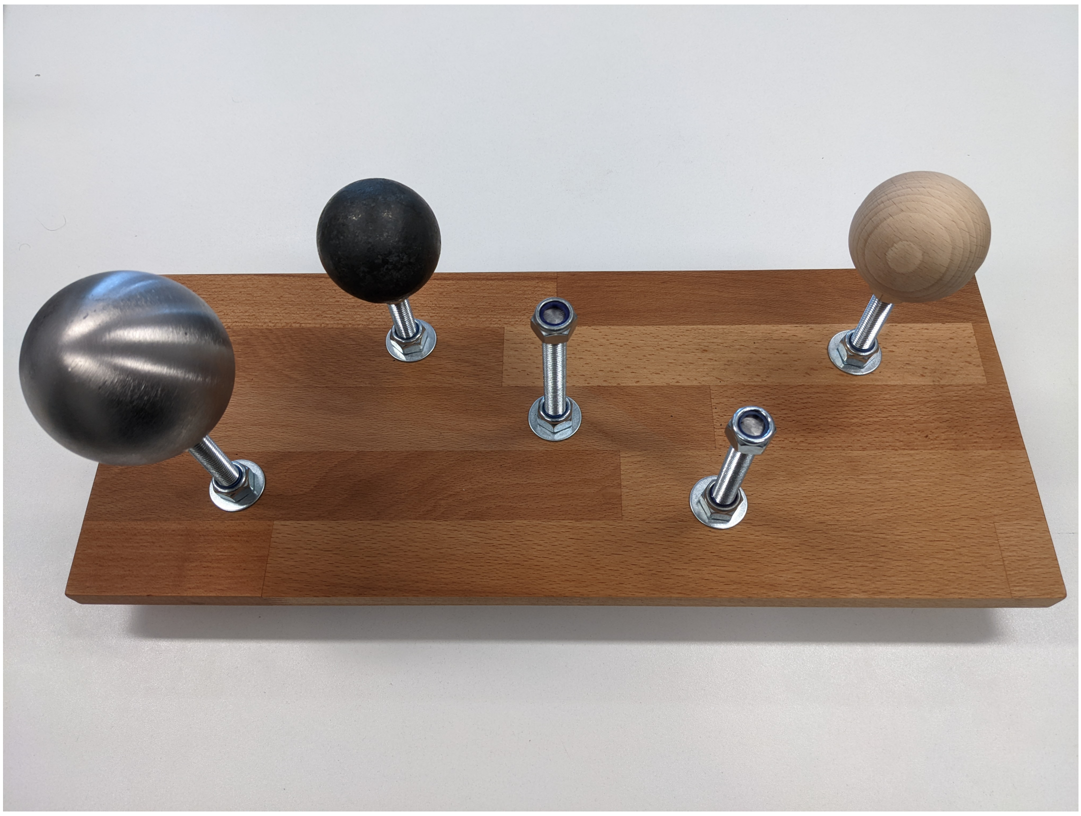


Figure 3.5 Constructed artefact for placement of spherical objects

3.4.2 Sphere spacing error E_s

Complementary to the sphere shape test, this test measures the error of relative centre-centre distances of the estimated sphere models. The reason for measuring this error is that the estimated radii may be close to the ground truth. However, the spheres in the scene may appear conically distorted due to an increasing angle of incidence. So if the estimated radius is correct, the position will be shifted.

Given set C_T of actual sphere central coordinates and C set of estimated central coordinates, the error is defined as follows.

$$E_s = \frac{1}{\binom{\#C}{2}} \sum_{i,j \in \{1..\#C\}, i<j} ||C_{Ti} - C_j|| \tag{3.2}$$

In other words, E_s is the mean absolute error between the actual distances and the estimated ones.

3.5 Tests on planar artefacts

This test suite measures the ability of each of the devices to estimate the depth and shape of planar surfaces with respect to the incidence angle of the artefact with the camera sensor and the distance of the artefact from the camera sensor.

The planar artefact should be positioned such that its centre is aligned with the centre of the camera sensor. The distance of the artefact to the camera and the incidence angle are measured from centre to centre.

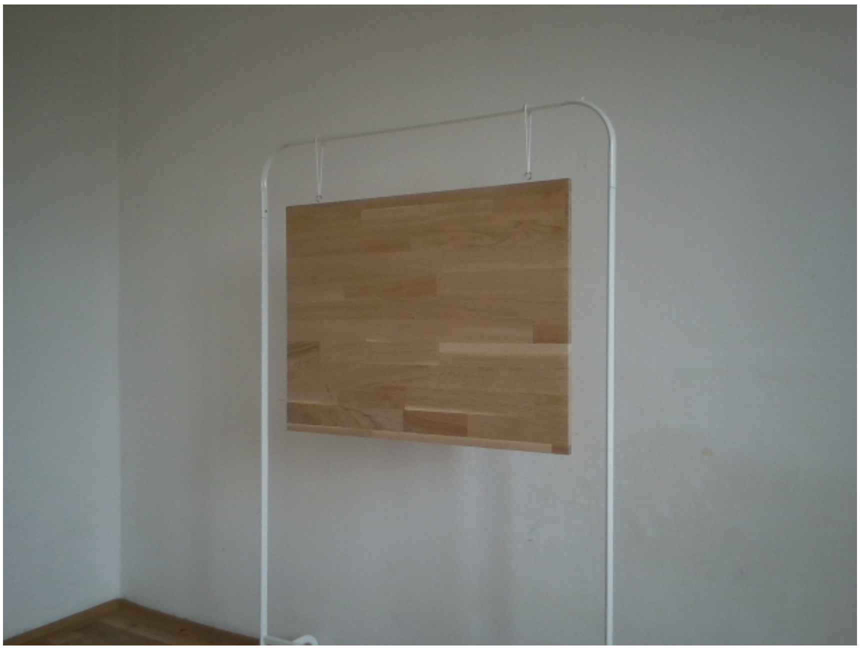


Figure 3.6 A wooden board hanged from supporting structure.

3.5.1 Distance estimation error E_d

Since the centre of the pictured artefact is aligned with the centre of the camera sensor, the distance estimation error is defined as the mean difference between the depth of all points within the ROI and the actual acquisition distance.

Let P be a 3D point cloud of the ROI and d the distance at which the point cloud was acquired. E_d is then defined as follows.

$$E_d = \frac{1}{\#P} \sum_{(x,y,z)\in P} (d - z) \tag{3.3}$$

The reconstructed 3D point cloud may contain empty areas because the camera sensor cannot estimate the distance. The mean distance will therefore be shifted and reflected in the measured error. Such empty areas may appear in areas with too high of an incidence angle with the sensor.

3.5.2 Flat-form distortion error E_f

Complementary to the distance estimation test, this test measures the distortion of the pictured planar artefact.

The observed error in this test is the mean absolute distance of a point cloud to a best-fit plane. A plane model is fit to a 3D point cloud depicting the face of the planar artefact, giving the following function for the z part of the coordinate derived from the scalar equation.

$$Z(x,y) = i + \alpha x + \beta y \tag{3.4}$$

Where i is the y-intercept of the plane, α and β are the scalars. With point cloud P of the region of interest, the error is defined as follows.

$$P_d = \frac{1}{\#P} \sum_{(x,y,z)\in P} ||(x,y,z) - (x,y,Z(x,y))|| \tag{3.5}$$

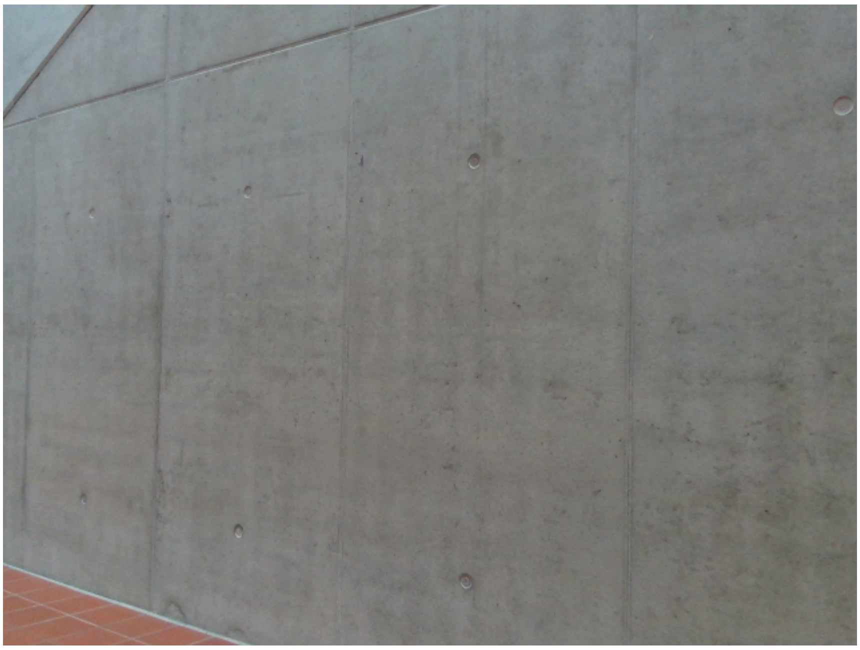


Figure 3.7 An exposed concrete wall

3.6 Concave surface profile estimation error E_p

This test compares the structural resolution of the devices. The purpose of this test is to determine the ability of each device to capture concave regions, i.e. to show the distortion in regions where a rapid change of depth occurs. Another point of interest here is to show how small a concave region can be registered with each device.

Let P be a point cloud of a concave region on the top face of the artefact obtained by scanning, and let P_T be a set of points depicting the real concave region on the top face of the artefact. P should be registered so that the distance to P_T is minimal, P_T should be positioned so that it is parallel to the xy plane.

For each point $p \in P$, let point $s \in P_T$ be a point satisfying the following criteria.

$$||(s_x, s_y) - (p_x, p_y)|| = \min\{d | d = ||(q_x, q_y) - (p_x, p_y)||, q \in P_T\} \tag{3.6}$$

The error of point p is then the L2 norm of $s - p$, and E_p is the mean value of all errors of points in P.

3.7 Multi-view fusion error E_m

Similar to the concave surface profile estimation error, this test also measures the structural resolution of each of the devices. The error is measured on a reconstructed model of a scanned object, therefore it requires a series of consecutive point clouds taken from different perspectives.

The point clouds are registered in a common coordinate system, and the background of the scenes is segmented out. The resulting scan is then registered on a reference 3d model, and the mean absolute error is calculated.

Due to the process of registration of successive point clouds, there is a certain degree of uncertainty due to the cumulative error of registration algorithms such as ICP and the implementation of the registration pipeline.

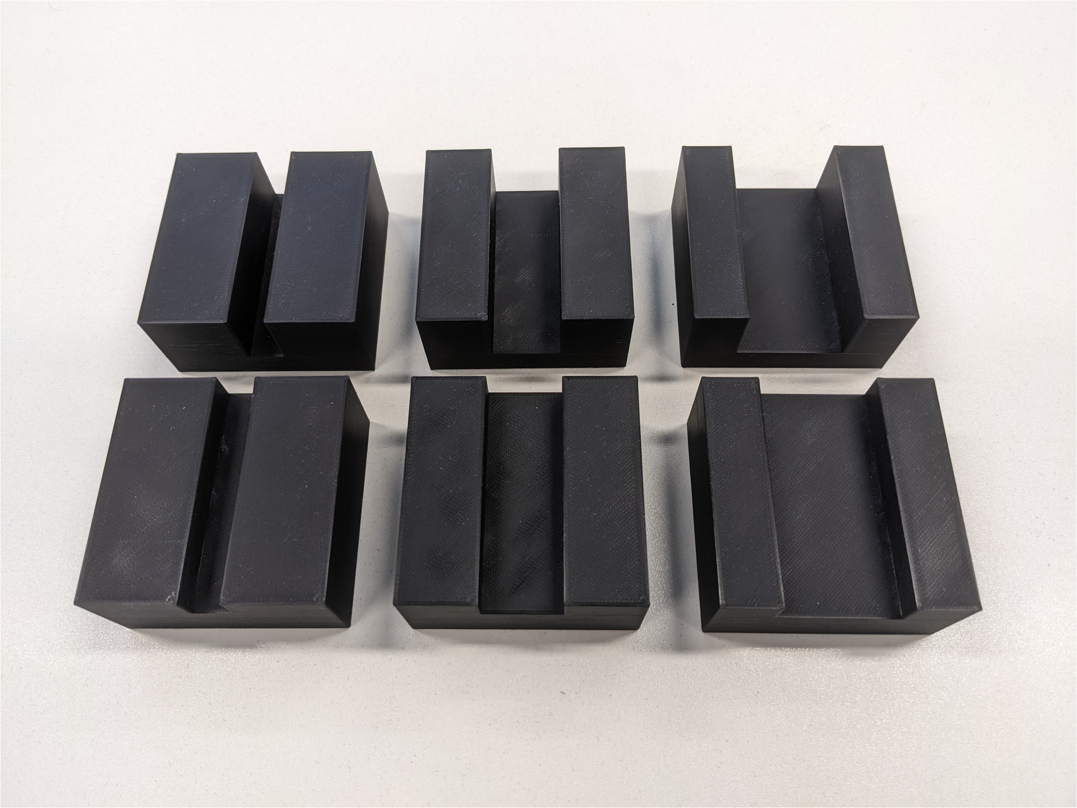


Figure 3.8 Cuboidal artefacts

The error is defined as follows. Let P be a point cloud of the scanned object, P_T a point cloud sampled from a triangle mesh model of the reference object and $n : \mathbb{R}^3 \to \mathbb{R}^3$ a function returning the nearest neighbor of point $p \in P$ from P_T.

$$E_m = \frac{1}{\#P} \sum_{p \in P} ||p - n(p)|| \tag{3.7}$$

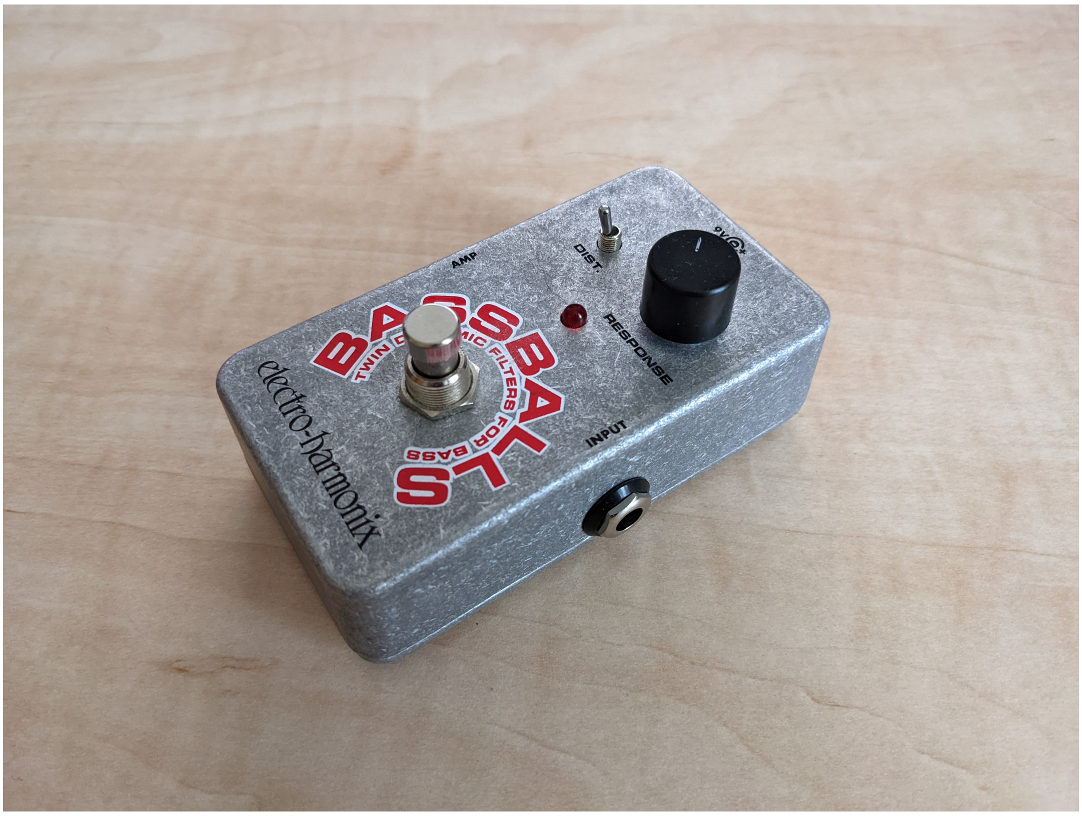


Figure 3.9 A free-form artefact

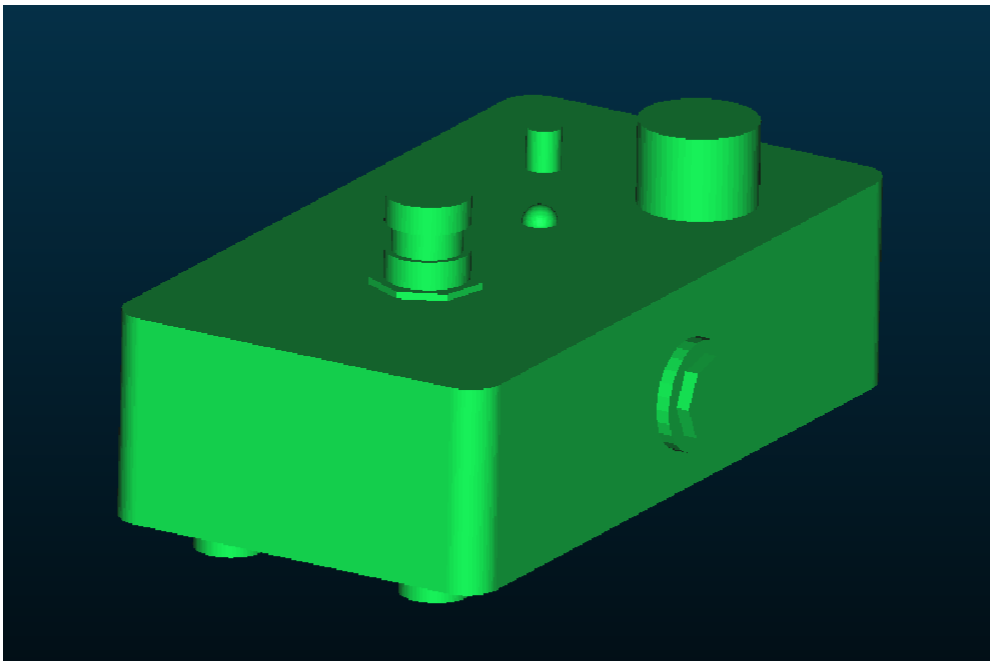


Figure 3.10 A 3D model of the free-form artefact

Chapter 4

Measurements

Every compared device was left running for 30 minutes before each measurement to reach thermal stability. The settings for each device were identical in every test. While this may not be the ideal setting for every application, it gives a reasonable basis for comparison. The acquisition distance was measured after the placement of each camera by a laser range finder.

4.1 Intel RealSense D415

The acquisition of data from this camera was done by a custom script written in Python, with the use of RealSense SDK Python language bindings.

Two different presets were compared, the High Density preset and the High Accuracy preset, to see both extremes of which the device is capable. The presets are referred to as RealSense D415 (A) and RealSense D415 (D) for High Accuracy and High Density, respectively.

In terms of acquisition pipeline settings, the spatial, hole filling and temporal filters were used with default parameters, no median filter for the blurring of the generated depth maps was used.

Both the RGB and depth images were used, the depth image was aligned with the RGB image, and the aligned images were then saved. The depth images were saved in the NumPy [1] matrix format (.npy) and the RGB images in the Portable Network Graphics format.

4.2 Basler ToF

Acquisition from this camera was done by the use of the Basler ToF Viewer program [29]. The program only allows the saving of ordered PCD files, normalized depth images and intensity and confidence maps. Only the intensity maps and point clouds were used.

To obtain non discretized depth image, the z-channel of the ordered PCD file was extracted into a depth image.

The temporal and spatial filters with default settings were used during the acquisition.

4.3 OAK-D-Lite

Both the High Accuracy and the High-Density settings were compared. The settings are referred to as OAK-D-Lite (A) and OAK-D-Lite (D) for High Accuracy and High Density, respectively.

[1] https://numpy.org/

Both temporal and spatial filters were used with the default settings, and no median filter for the blurring of the generated depth maps was used.

The acquisition of RGB and depth frames was done by a custom script utilizing the DepthAI SDK. The depth frames were saved in the NumPy matrix format, and the RGB frames were saved in the Portable Network Graphics format.

4.4 Measurement of spherical artefacts

Objects described in section 3.2.1 were used as measuring artefacts.

A total of 60 images were captured in sets of 4. Each set contains images of different spheres shot with increasing distance.

The sphere holding artefact was placed on a camera testbed and was then then pictured from the distances of 500, 600, 700, and 800 mm, measured from the top of the testbed to the camera sensor.

Three different sets of spheres were used, the first set consists of 5 wrought iron spheres with a 30 mm radius, the second set consists of 4 wrought iron spheres and one aluminium sphere with a 40 mm radius and the third set consists of 5 wooden spheres with a 30 mm radius.

Therefore a total of 12 images were shot by each camera.

The spheres were segmented out from the background by colour for the RGB-D cameras and by depth for the Basler ToF. Figures 4.1 and 4.3 show unmasked colour and depth images. Figures 4.2 and 4.4 show colour and depth images after applying a mask.

The reconstructed point clouds were then fitted with 3D sphere models.

4.4.1 Sphere shape error E_r

Using the fitted models of the spheres visible in the shot, a mean absolute deviation of estimated radii from the ground truth was calculated. Figure 4.5 shows the reconstructed spheres with best-fit sphere models.

4.4.2 Sphere spacing error E_s

Using the fitted models of the spheres visible, a mean absolute deviation of the distances of spheres visible in the shot was calculated.

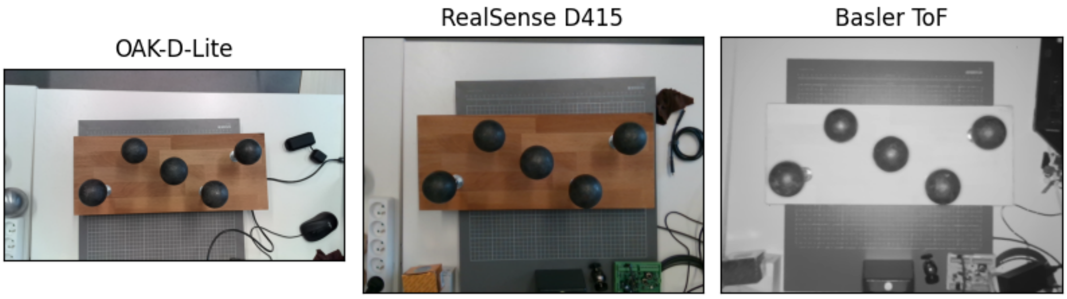


Figure 4.1 Wrought iron spheres shot from the distance of 700 mm.

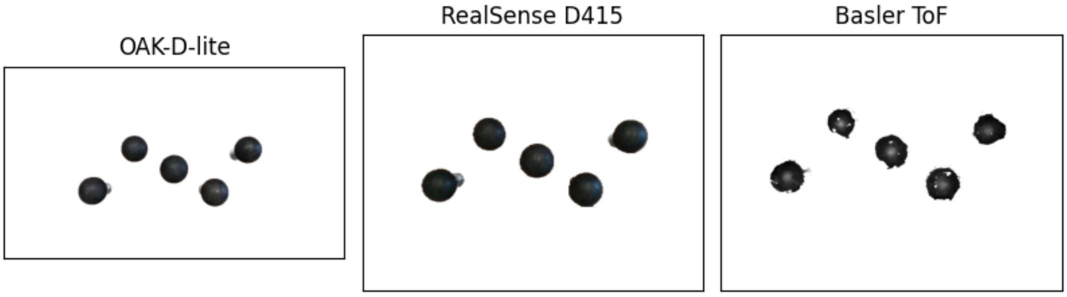


Figure 4.2 Masked wrought iron spheres shot from the distance of 700 mm.

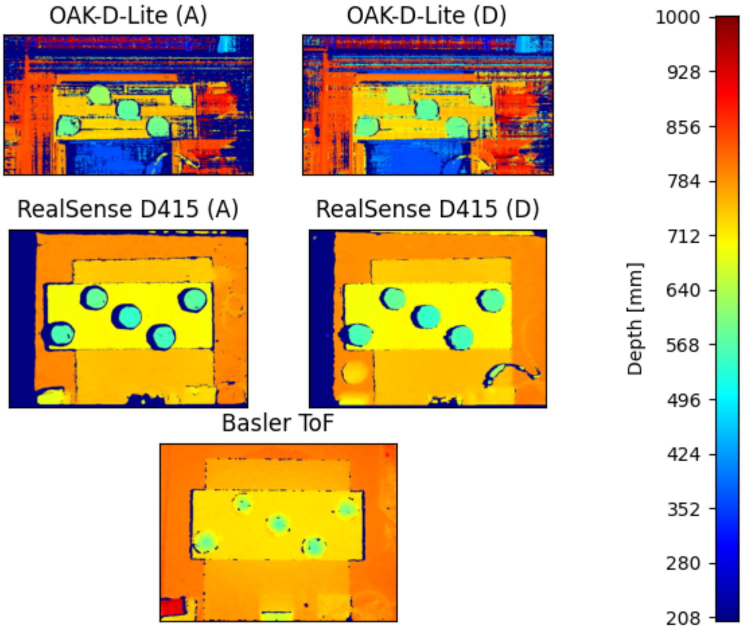


Figure 4.3 Depth maps of wrought iron spheres shot from the distance of 700 mm.

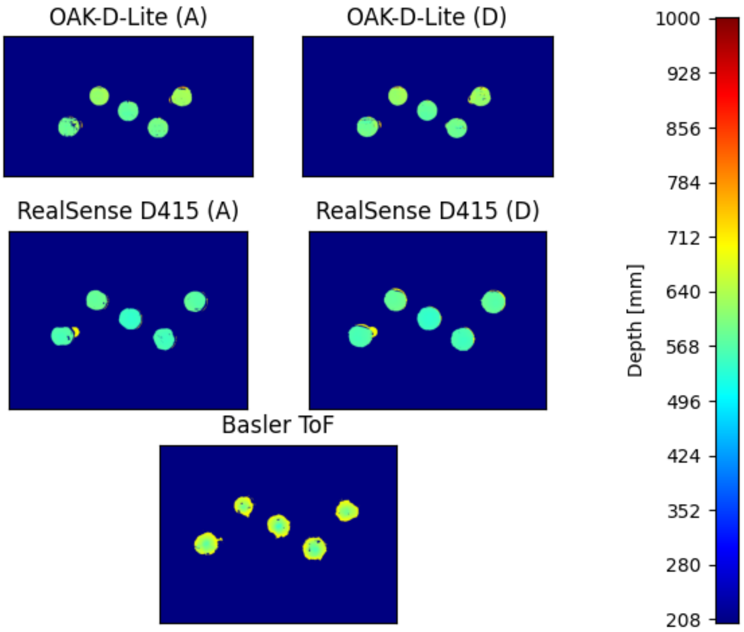


Figure 4.4 Masked depth maps of wrought iron spheres shot from the distance of 700 mm.

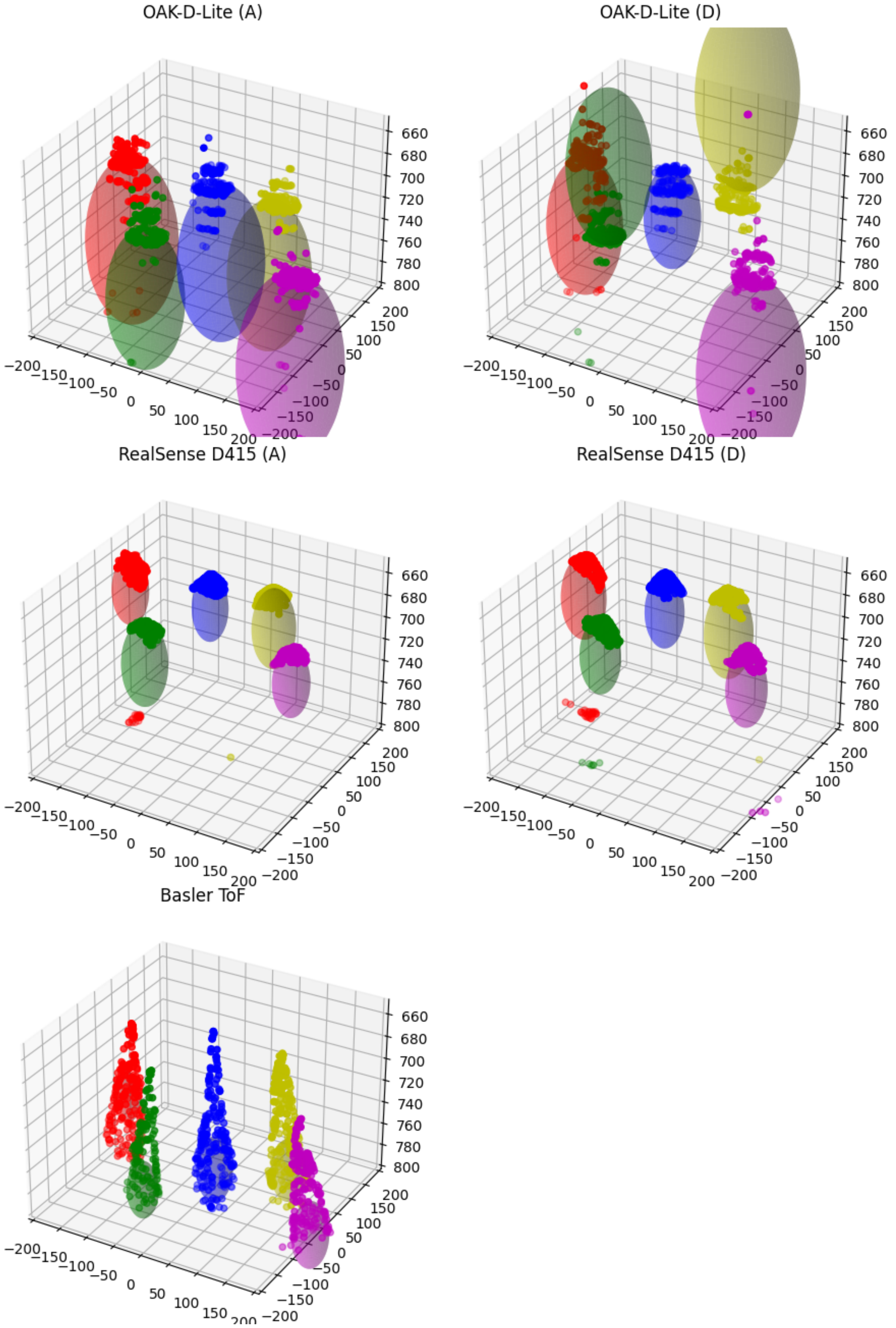


Figure 4.5 Reconstructed 3d spheres with fitted sphere models

4.5 Measurement of planar artefacts

Objects described in section 3.2.2 were used as measuring artefacts. 180 images were captured, 90 of which were of a wooden board and 90 of an exposed concrete wall. Each set was shot at the distance of 500-3000 mm with 500 mm increments at an angle of 30, 60, or 90 degrees and contains six images. Therefore a total of 18 images were shot by each camera in each scene. An example of captured images and depth maps can be seen in figure 4.6 and figure 4.8.

For the two RGB-D cameras, OAK-D-Lite and RealSense D415, the ROI from the board dataset was segmented out in the HSV colour space from the coloured image, and the generated mask was used to mask out the uninteresting regions in the depth image. For the Basler ToF camera, a combination of segmenting by depth and intensity was used.

In order to obtain a segmented ROI in the concrete wall dataset, the respective masks generated in the board dataset were used, since the acquisition distances and angles were identical.

An example of segmented colour images and depth maps can be seen in figure 4.7 and figure 4.9 respectively.

The segmented depth maps were then projected back into 3D space. An example of the reconstructed point clouds of a planar artefact is displayed in figure 4.10.

4.5.1 Distance estimation error E_d

Using the reconstructed model of the planar surface and real measured distance, a mean of the depth estimation error of the points was calculated. Figure 4.11 shows a 2D side view of the generated point clouds together with a marker for ground truth.

4.5.2 Flat-form distortion error E_f

Using the reconstructed model of the planar surface, a plane model was fitted into the 3D points using RANSAC. Example of such model can be seen in figure 4.12. Then the mean absolute error and variation of the measured 3D points from the fitted model were calculated.

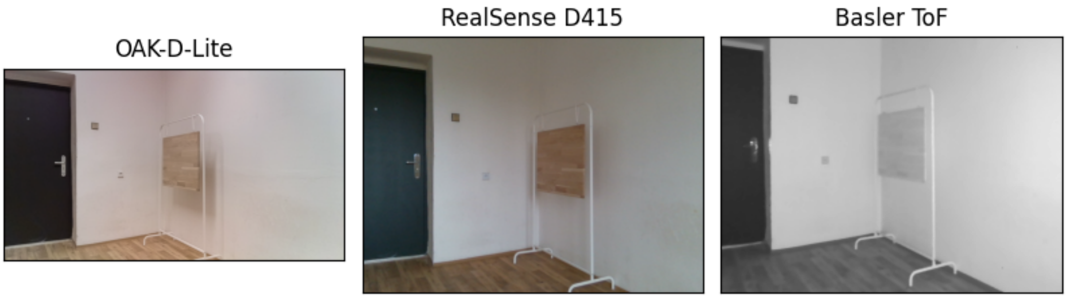


Figure 4.6 Wooden board captured from 3000 mm under a 60-degree incidence angle.

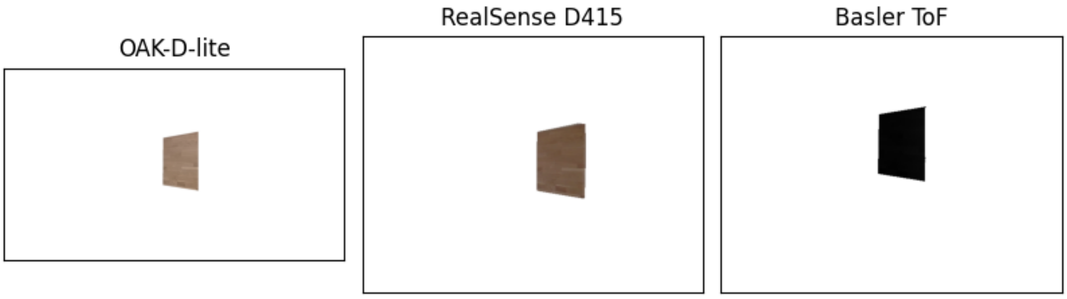


Figure 4.7 Masked wooden board captured from 3000 mm under a 60-degree incidence angle.

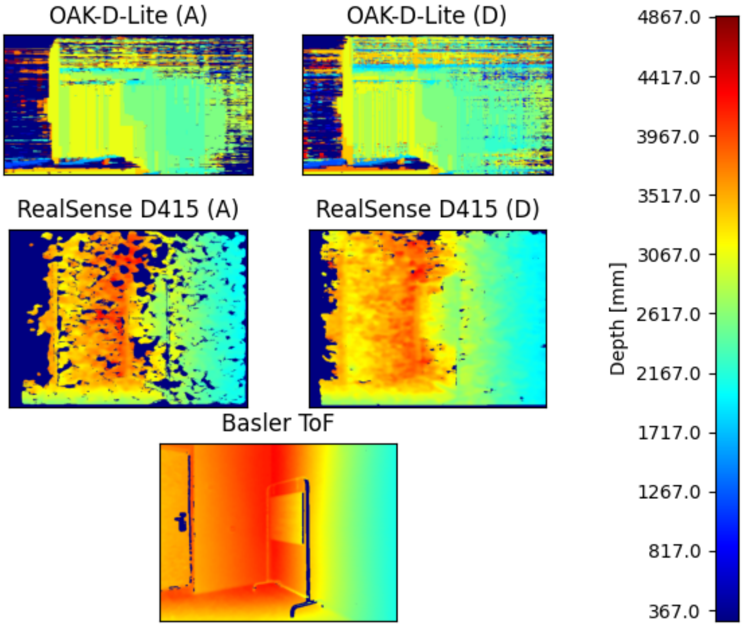


Figure 4.8 Wooden board depth maps captured from 3000 mm under a 60-degree incidence angle.

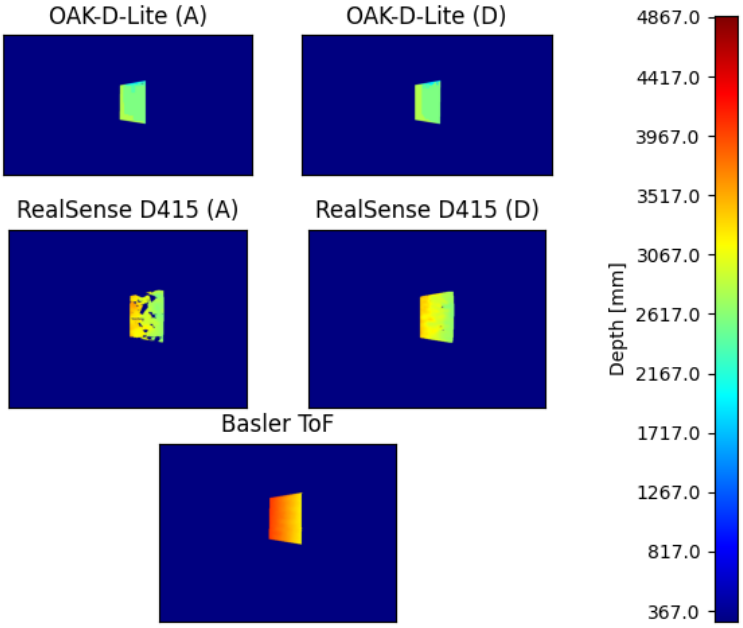


Figure 4.9 Masked wooden board depth maps captured from 3000 mm under a 60-degree incidence angle.

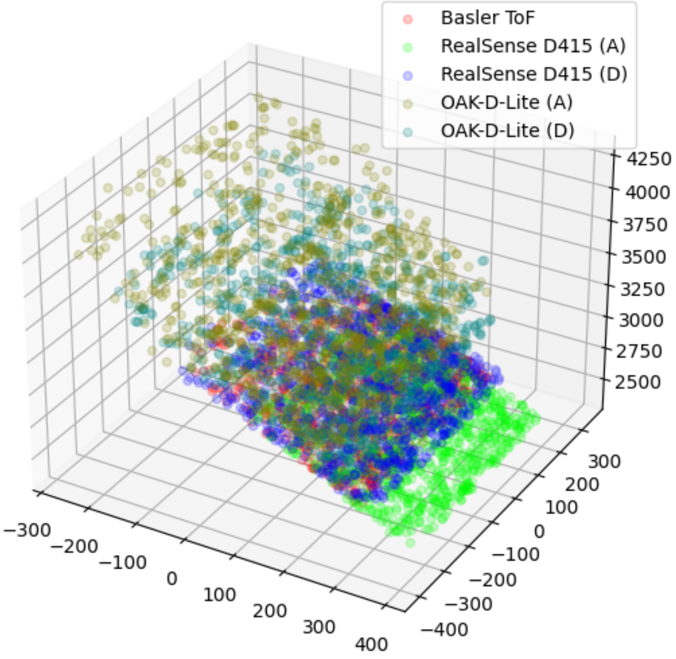


Figure 4.10 Example of reconstructed point clouds of a wooden board

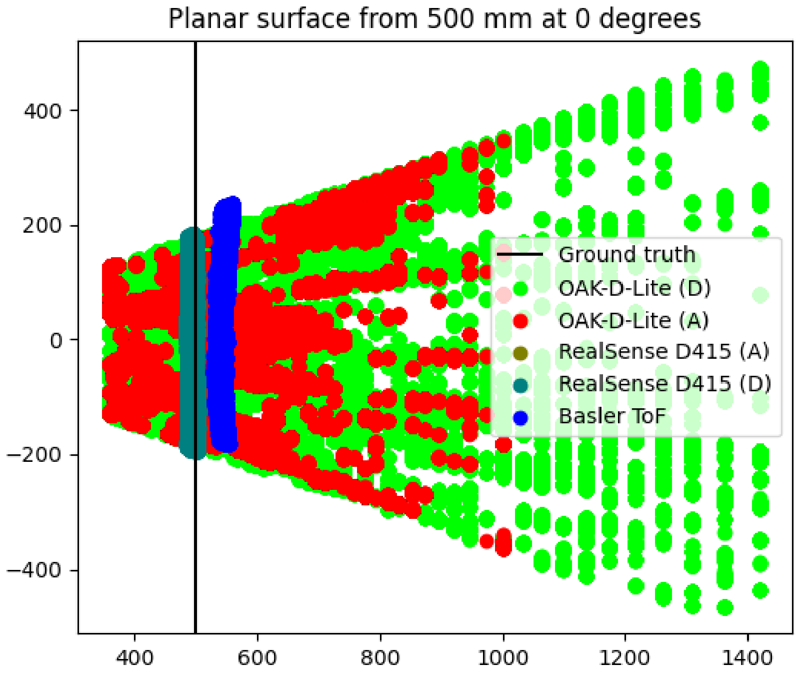


Figure 4.11 Side view of wooden board point clouds generated by different sensors, projected to 2D.

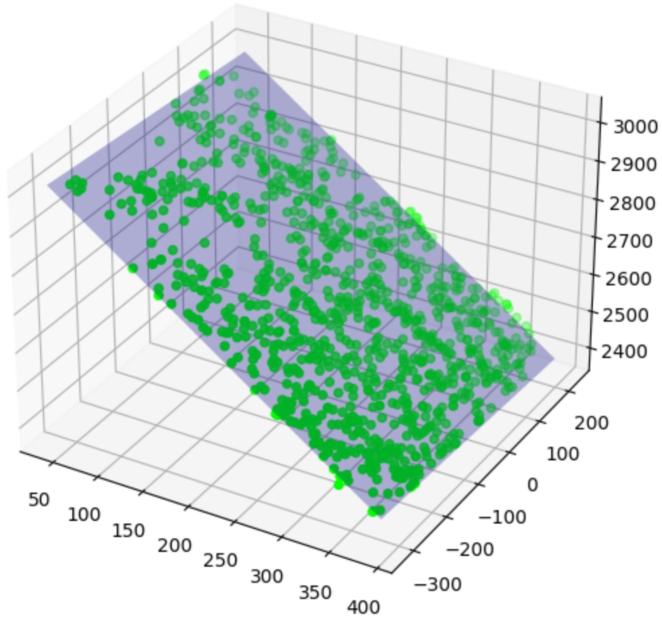


Figure 4.12 A point cloud of a wooden board with a best-fit plane model

4.6 Measurement of concave surface profile

Each of the six cuboidal artefacts from section 3.2.3 were placed on the test-bed under the camera sensor so that the artefact and the camera sensor are parallel and pictured from distances of 500, 600, 700, and 800 mm, measured from the top of the testbed to the camera sensor. Therefore 24 pictures were captured by each camera, 4 of each cuboidal artefact.

Similarly, as in other tests, the ROI was segmented out of the scene by either colour or depth and projected back into 3D space. Figures 4.13, 4.14 and 4.15 show the colour images of the scene, depth maps of the scene and depth maps of the top profiles generated by the cameras.

The outliers on the edge of the ROI were removed in order to obtain a point cloud representation of the top face of the cuboidal artefact. In the next step, the 3D polygon model of the cuboidal artefact was positioned so that its bottom face was parallel with the xy plane. Then it was sampled to yield a point cloud representation of the model, and the top face was segmented out. The two-point clouds were then aligned with the use of ICP.

From both of the point clouds, only the concave region was used for the calculation of the error, so that the top face would not skew the result, and the deviation from ground truth would be more apparent.

Correspondences between the two-point clouds were obtained by constructing a k-d tree from the model point cloud, retrieving the nearest neighbours for each scanned point in a radius of two times the cut depth and picking the nearest point in the x and y dimension.

The mean distance and variation of the measured profile from the model were then calculated. Figure 4.16 shows the reconstructed top profiles with inliers coloured green and outliers red. Points classified as inliers have an error lower or equal to the mean error, while points classified as outliers have an error larger than the mean. Similarly, figure 4.17 shows the side view of the measured profiles along with the profile of the 3D model.

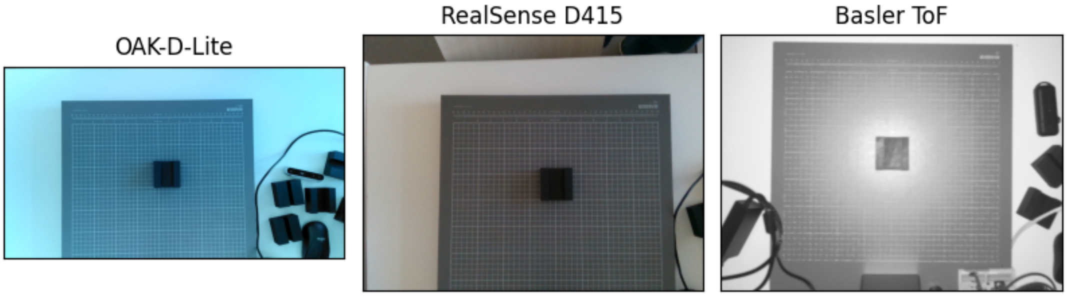


Figure 4.13 Colour images of the cuboidal artefacts scene

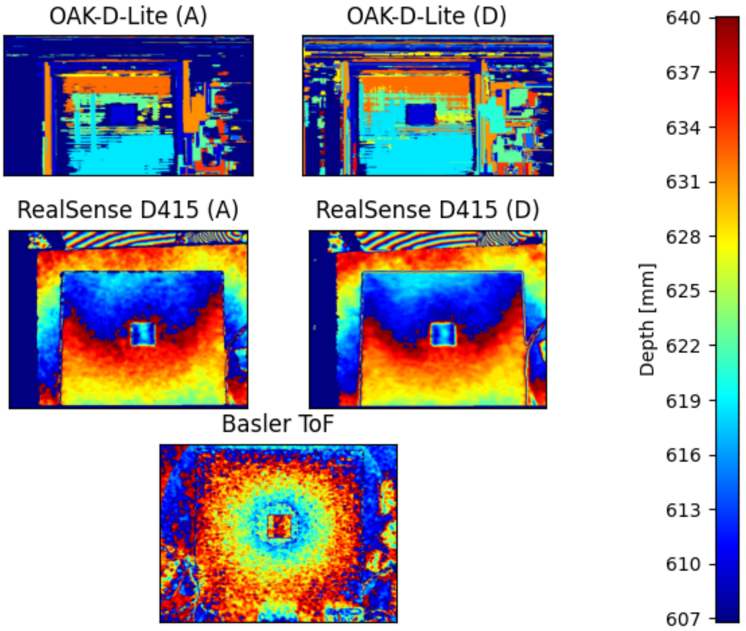


Figure 4.14 Depth maps of the cuboidal artefacts scene

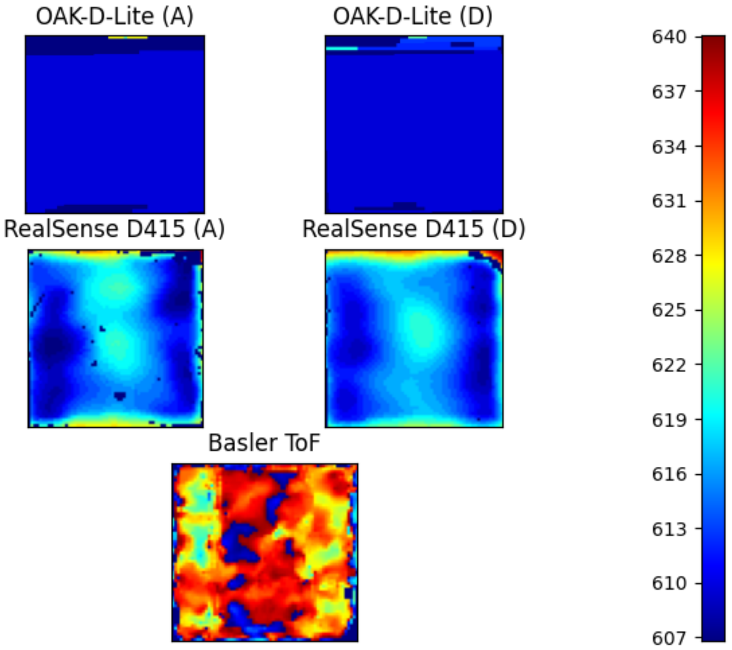


Figure 4.15 Depth maps of the top faces of the artefacts

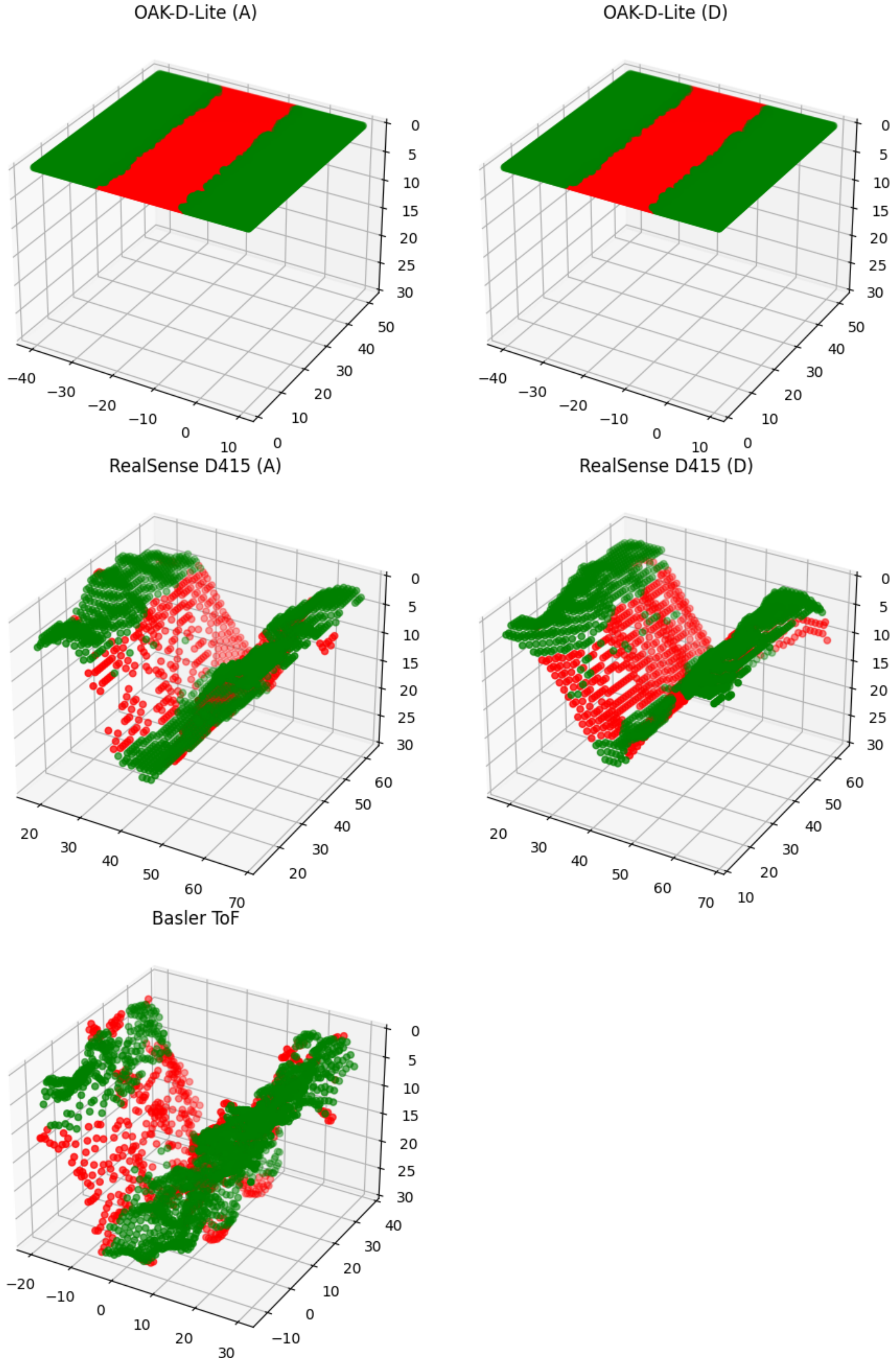


Figure 4.16 3D scene of the profiles generated by the cameras with coloured inliers and outliers.

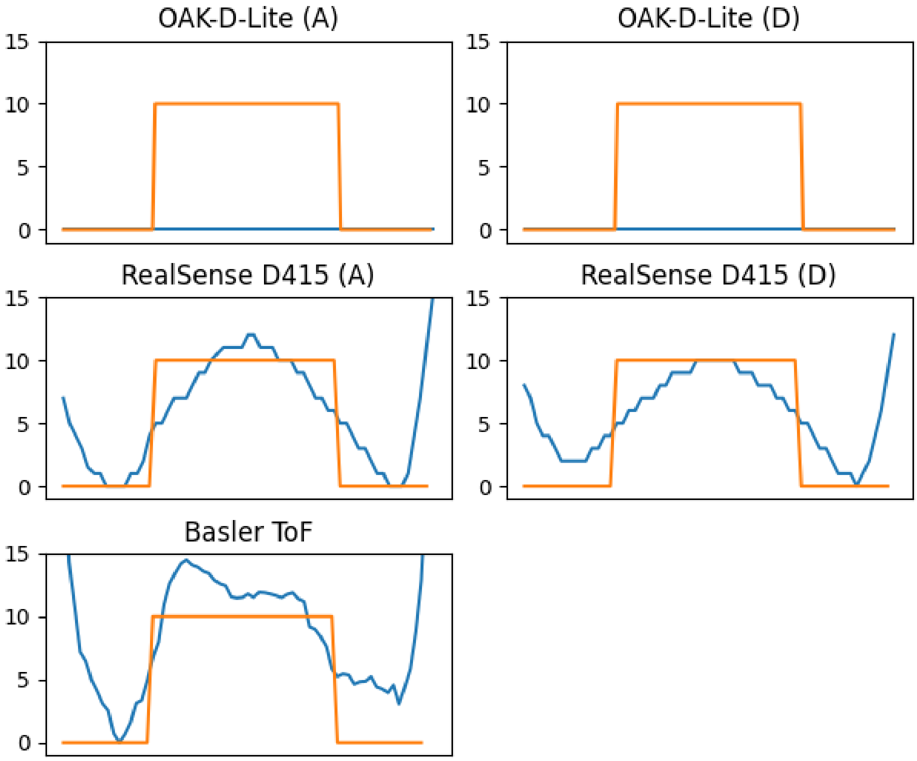


Figure 4.17 Side view of the mean distances along the profile with ground truth

4.7 Multi-view measurement

All of the devices were tested, but the complete measurement was completed only with the RealSense D415 in both the High Accuracy and the High Density settings.

Basler ToF showed a high amount of distortion due to the high incidence angles and the glossy finish of the artefact. The distortion is shown on figure 4.18. Due to this distortion, ICP would not be able to register the point clouds correctly because the transformation generated by ICP is rigid.

OAK-D-Lite was also unable to produce sufficiently accurate depth maps to measure any meaningful error. An example of the depth maps containing a large amount of noise is displayed in figure 4.19.

The measurement with the RealSense D415 proceeded as follows. A series of 100 consecutive RGB-D images were taken from a distance of around 500 mm. The camera trajectory was similar in both measurements. An approximation of the trajectory is shown in figure 4.20. A precise trajectory could not be generated from the series by the use of RGB-D odometry, as the scene was primarily flat with the artefact situated in the centre and lacked enough distinct key points.

The consecutive point clouds were registered by the ICP algorithm. The ICP was not initialized by any precomputed initial transformation, as the distance between the consecutive point clouds is minimal. Only points not already present in the scan were added to the resulting scan during the run of the fusion pipeline. The pseudo-code listing 4.1 summarizes the entire process of the registration and fusion of the successive point clouds. The resulting scans are shown in figure 4.21.

Code listing 4.1 Pseudocode of point cloud fusion algorithm

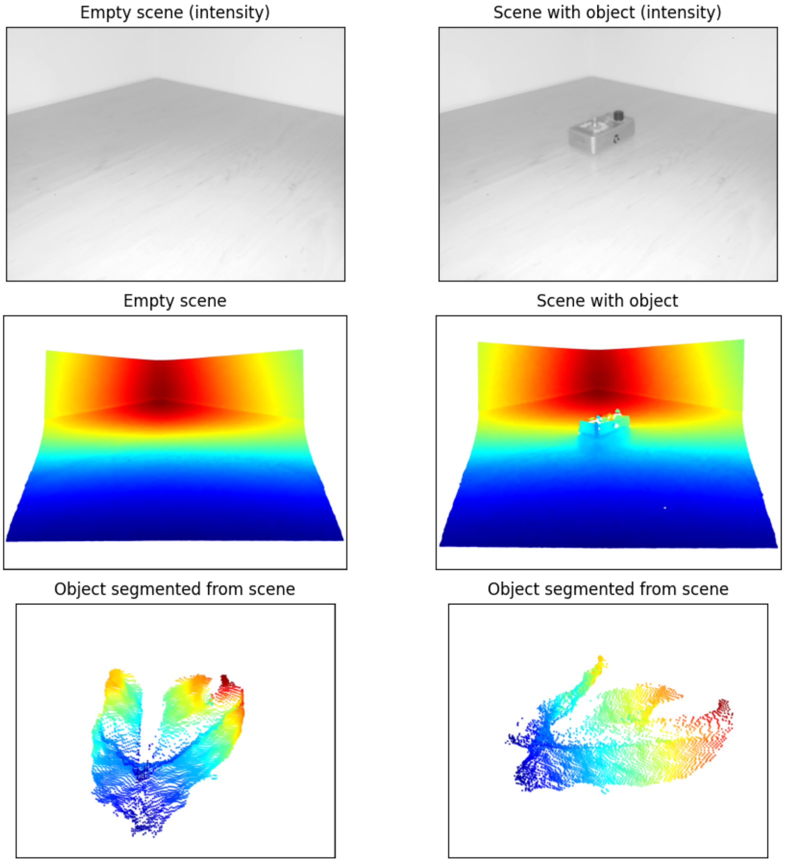
```
def fusion(pcd_it, lower_bound, upper_bound):
        scan = pcd_it.grab_next_frame()
                .reconstruct_pc()
                .downsample()

        scan = segment_roi(scan)

        while (next_frame = pcd_it.grab_next_frame()).is_valid():
                next_frame = next_frame.reconstruct_pc().downsample()
                next_frame = segment_roi(next_frame)
                transformation = ICP(scan, next_frame)
                scan.transform(transformation)
                distances = next_frame.calculate_distance(scan)

                scan.add_points(
                        next_frame[lower_bound < distances < upper_bound]
                )

        return scan
```



Empty scene (intensity)

Scene with object (intensity)

Empty scene

Scene with object

Object segmented from scene

Object segmented from scene

Figure 4.18 Basler ToF distortion of the free-form artefact

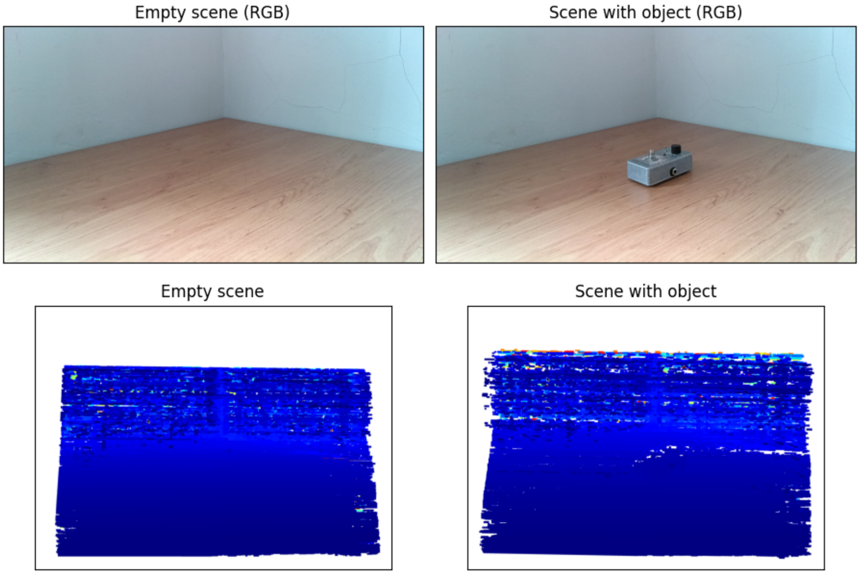


Figure 4.19 OAK-D-Lite point cloud noise in a free form artefact scene

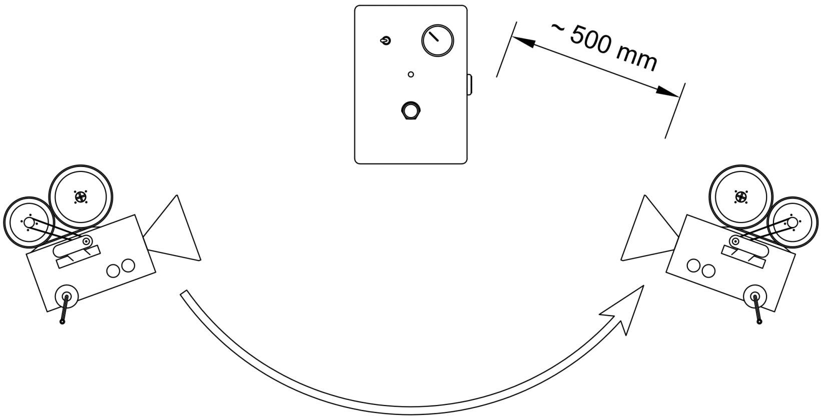


Figure 4.20 An approximate trajectory of the RealSense D415 during the measurement of the free-form artefact

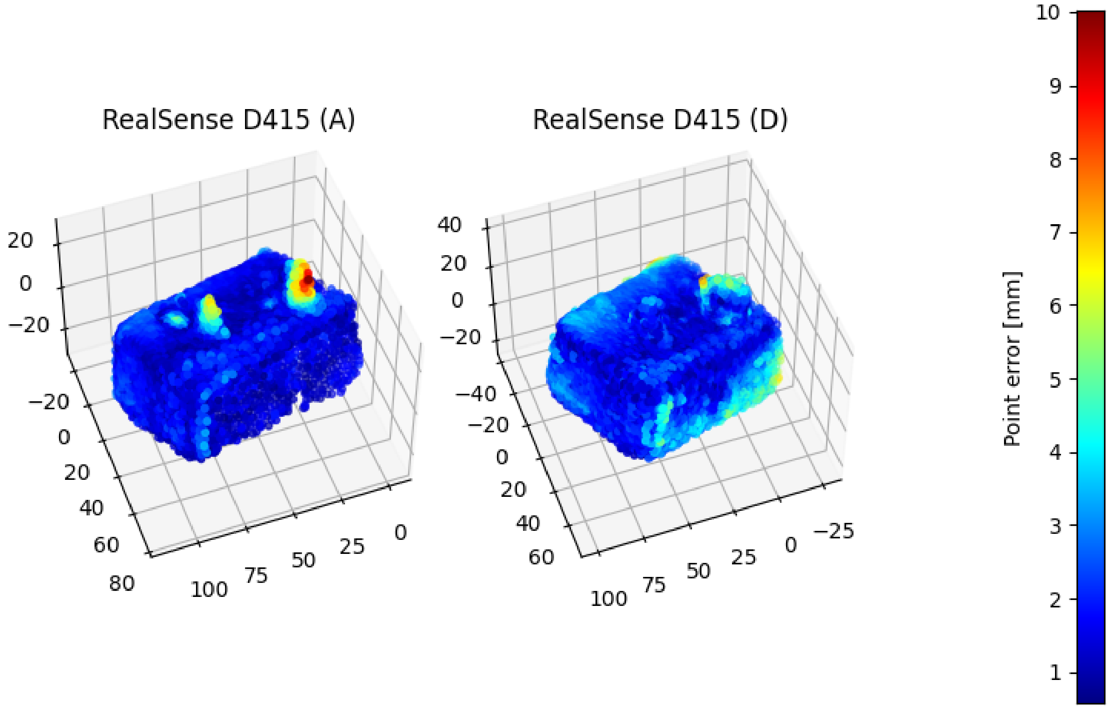


Figure 4.21 Reconstructed free-form artefact with a point error scale

Chapter 5

Results

The following values were measured in the captured datasets. Influences of measurement factors such as an increasing distance, an angle of incidence, and material surface characteristics are compared.

5.1 Spherical artefact shape and spacing estimation

Figure 5.1 shows the sphere shape error E_r, sphere spacing error E_s, and the mean count of measured points of the spheres depending on the measurement distance.

The OAK-D-Lite camera does not show any systematic error in radius estimation or spacing in any of the three datasets. As can be seen in figure 4.5, the reconstructed data is too noisy and does not resemble a top of a hemisphere. Due to the lack of resemblance, it is impossible to estimate a sphere model's parameters with RANSAC accurately.

On the two datasets containing metal spheres with shiny surfaces, the Basler ToF camera shows a substantial amount of conical distortion, hinted by the significant spacing error. The conical distortion is also visible in figure 4.5. The radius estimation error of Basler ToF seems to be lessened by an increasing measurement distance, most likely due to a lower intensity of the reflection. On the dataset with matte wooden spheres, the error is not as substantial, and the camera can capture the spheres fairly well.

The RealSense D415 was able to estimate the sphere radius and spacing with good accuracy and was not affected by the shiny or matte surfaces. No systematic error due to an increasing distance seems to be present in the measurement range.

5.2 Flat-form distortion

Figures 5.2 and 5.3 show the flat-form distortion error measured on a wooden board and a concrete wall respectively. The mean flat-form distortion error E_f is marked with a cross.

5.2.1 Wooden board dataset

With an incidence angle of 0 degrees, i.e. the sensor and the measured planar object in parallel, the Basler ToF shows the least amount of flatness distortion. The RealSense D415 performs similarly to Basler ToF on lower distances. However, it shows more distortion gradually with an increasing distance on both of the presets. The images obtained from OAK-D-Lite contain considerable noise when taken at lower distances. With an increasing measurement distance, the

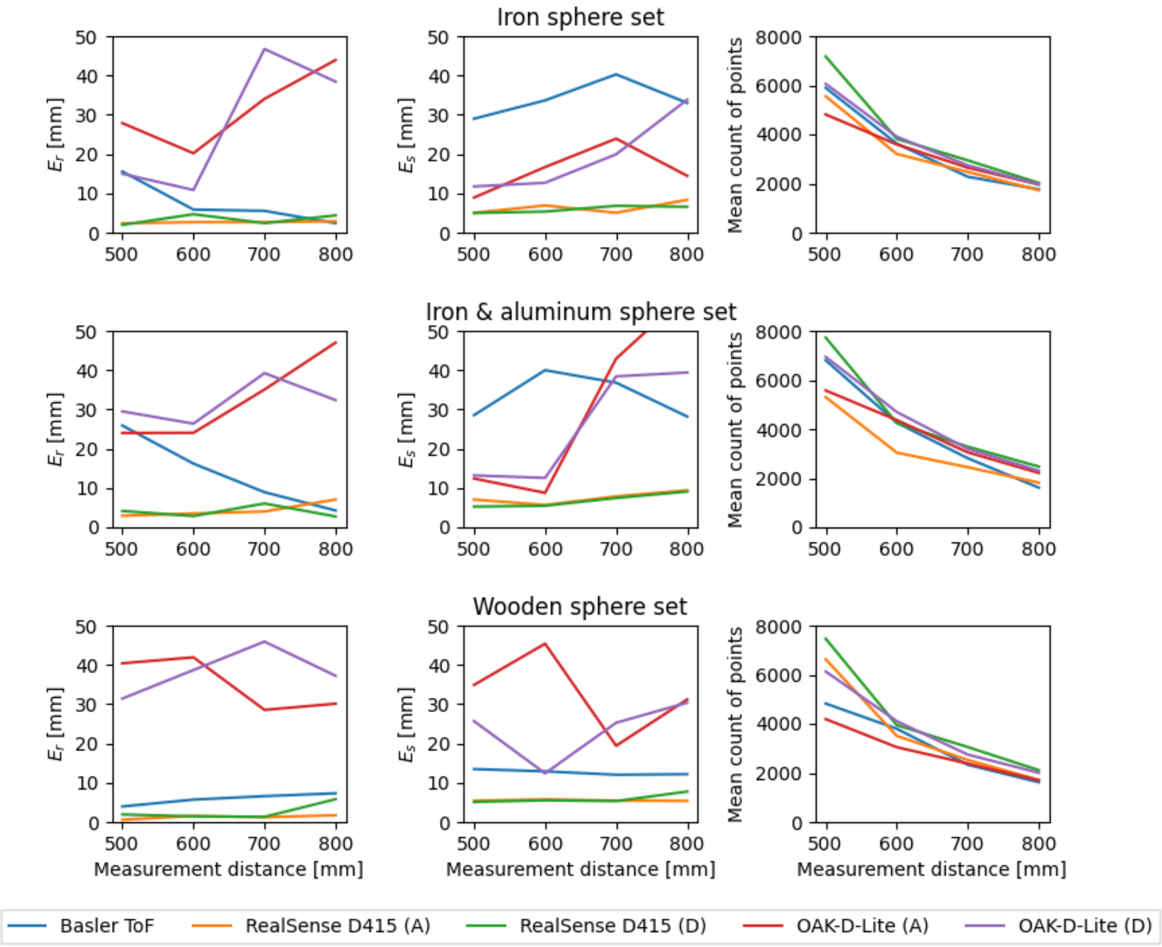


Figure 5.1 E_r, E_s, and mean count of points measured on spherical artefacts.

distance of the wooden board is estimated uniformly by OAK-D-Lite and is therefore completely flat. However, the box plot shows a non-zero mean value due to the reconstructed point cloud containing a low amount of outliers. The uniform depth estimation is most likely caused by a low number of visually distinct key points on the board for disparity matching.

Both the RealSense D415 and Basler ToF show a large amount of distortion, which seems to be caused by a combination of the increasing incidence angle and the surface characteristics of wood. Again, they both perform similarly on lower distances. With an increasing angle of incidence, the OAK-D-Lite can no longer estimate the depth uniformly and has the most significant flat-form distortion error.

5.2.2 Concrete wall dataset

In contrast to the wooden board, the concrete wall offers more visually distinct key points and it, therefore, should be better suited for passive stereo vision.

Both the Basler ToF and the RealSense D415 perform far better on the concrete wall than

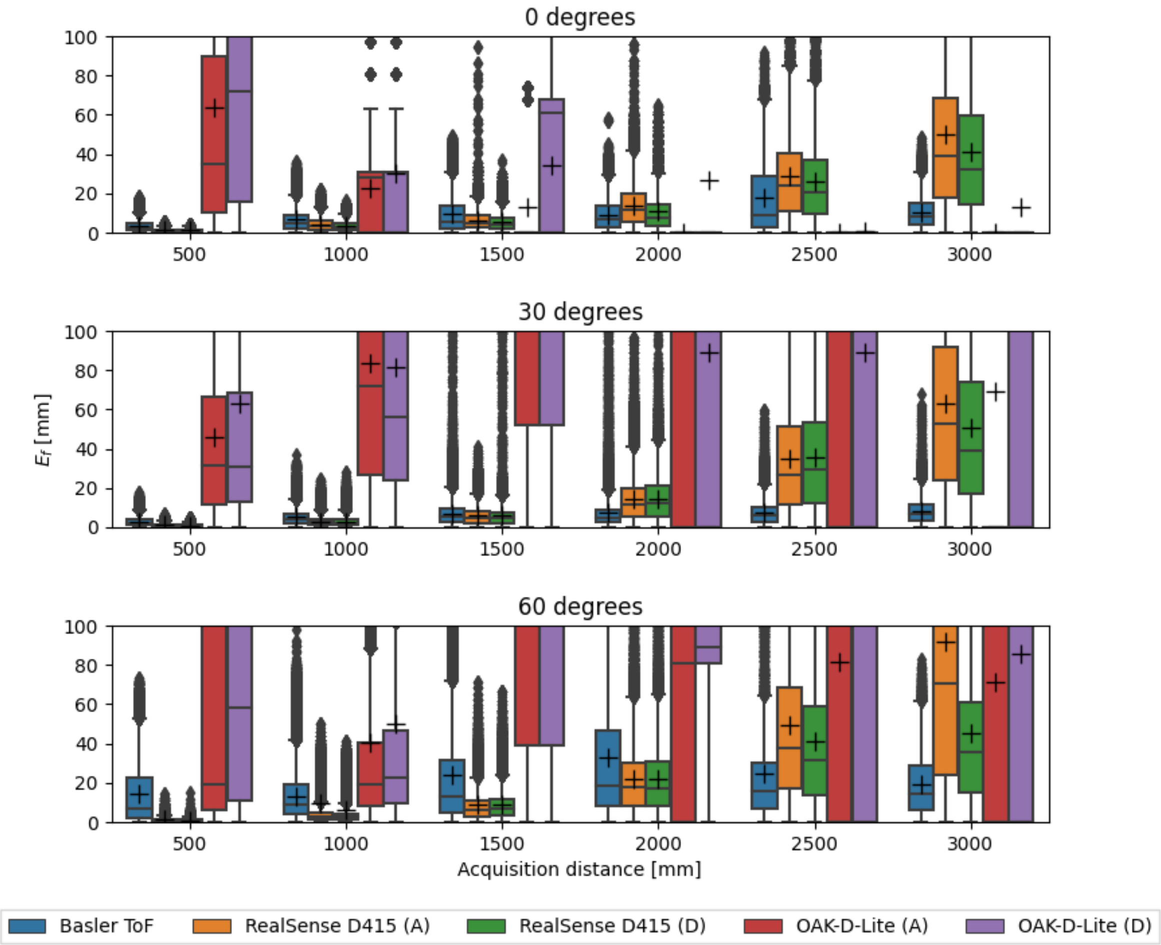


Figure 5.2 Flatness error measured on a wooden board dataset

they did on the wooden board. The exact reason for this is not immediately apparent. The surface texture and different reflectivity in the IR spectrum may both be influencing factors. It is, however, not possible to say for sure without a further investigation.

Due to the higher amount of distinct visual key points on the concrete wall, the OAK-D-Lite can perform much better and is comparable to the other cameras in the low range. The slight noisiness of the data generated by the OAK-D-Lite may be alleviated by using a mean filter or a similar blurring method.

5.3 Distance estimation

Figures 5.4 and 5.6 show the distance estimation error E_d measured on a wooden board and a concrete wall respectively.

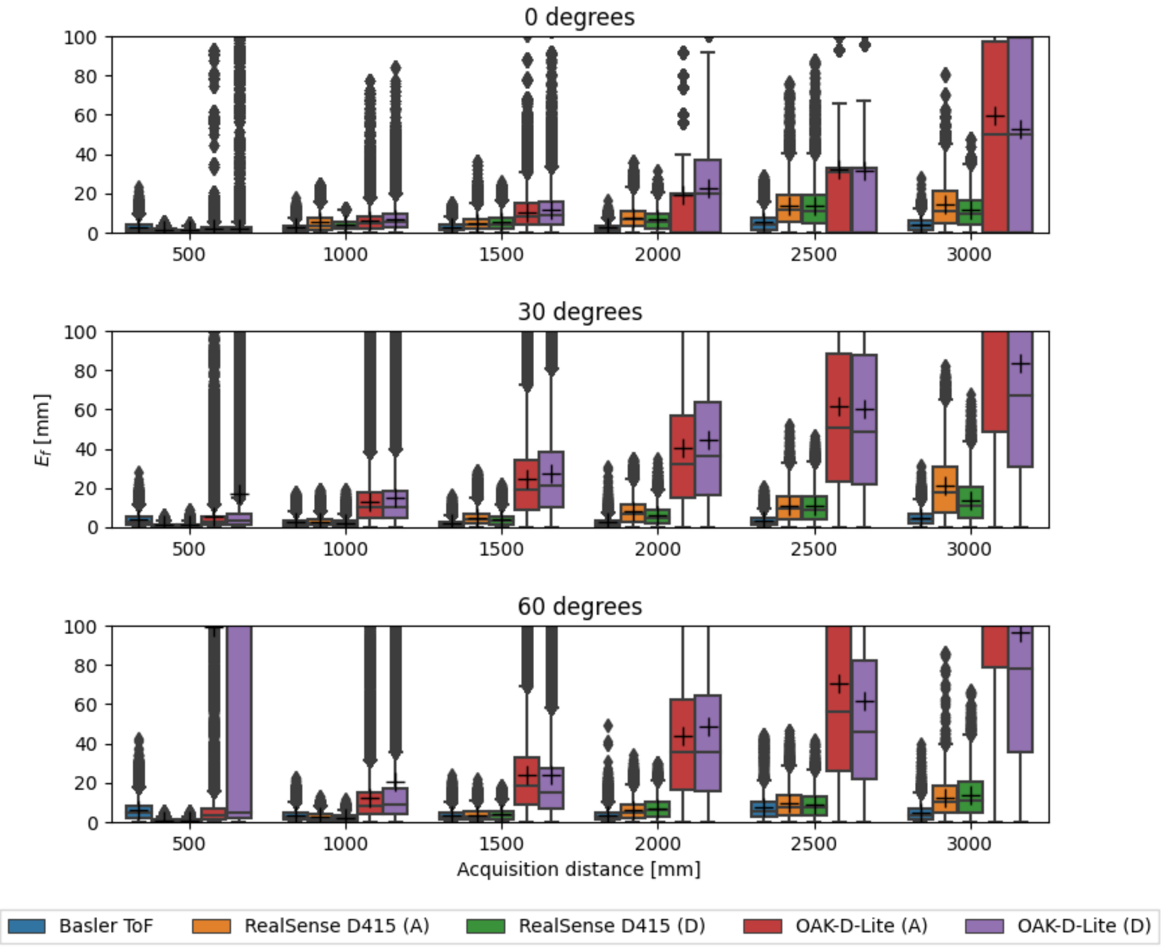


Figure 5.3 Flatness error measured on a concrete wall dataset

5.3.1 Wooden board dataset

All tested devices measure the mean distance of the wooden board reasonably well under a zero degree incidence angle in the short-range.

The Basler ToF starts showing quite a substantial deviation from the ground truth at around 1000 mm, and it systematically overestimates the distance of the wooden board with increasing measurement distance. The overestimation is far higher when a high incidence angle is introduced, reaching around 700 mm when measured under a 60-degree incidence angle at 3000 mm.

In contrast, the RealSense D415 can accurately measure the mean board distance.

The OAK-D-Lite performs similarly to the other cameras when the measurement angle is zero. The reason for this is the same as in the measurement of flat-form distortion error. With an increasing incidence angle and distance, it slightly underestimates the distance.

Figure 5.5 shows the amount of points captured on each picture. Since it doesn't hint at any large blind spots in the depth maps of any of the devices other than a slightly lower amount of points taken by the OAK-D-Lite at the 500 and 1000 mm measurement distances at an

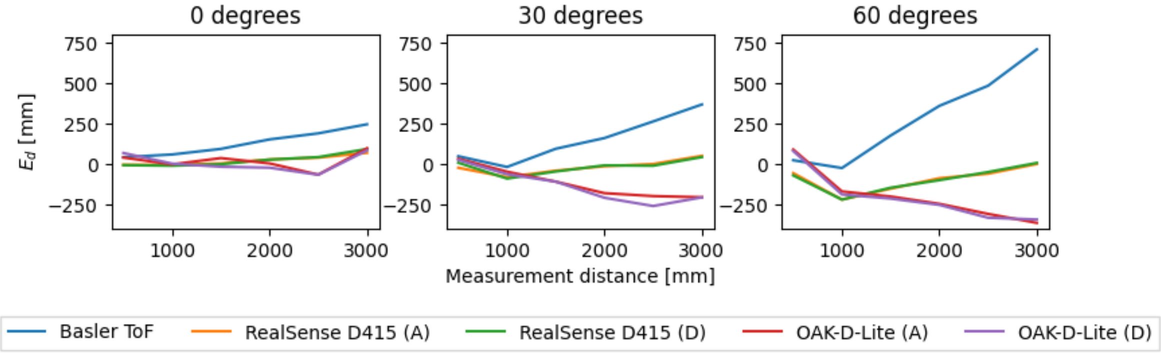


Figure 5.4 Distance estimation error measured on a wooden board dataset

incidence angle of 60 degrees, due to the ROI being too close to fit in the view. The reason for the underestimation and overestimation of the distances is therefore not obvious. The inaccuracy of the passive OAK-D-Lite is not surprising. However, the inaccuracy of the Basler ToF seems to be worthy of further examination.

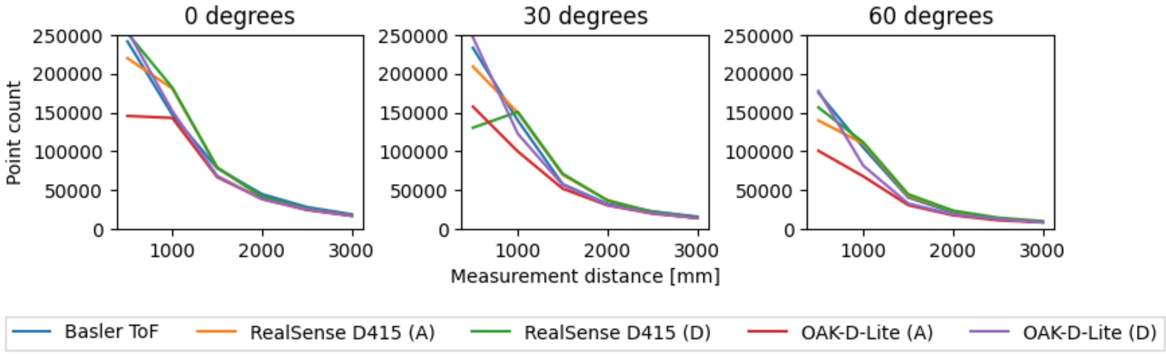


Figure 5.5 Mean amount of points captured on a wooden board.

5.3.2 Concrete wall dataset

The Basler ToF camera and the RealSense D415 were both able to measure the mean distance to the wall with reasonable accuracy under 0 and 30-degree incidence angles. However, they both showed an increased error in measurements under a 60-degree incidence angle.

The OAK-D-Lite shows a significant systematic distance measurement error under any incidence angle. The distance error is around 10% to 25% of the actual acquisition distance in the 2000 mm to 3000 mm range.

Figure 5.7 again hints at no large blind spots in the depth maps, other than for the OAK-D-Lite at 500 and 1000 mm at 60 degrees.

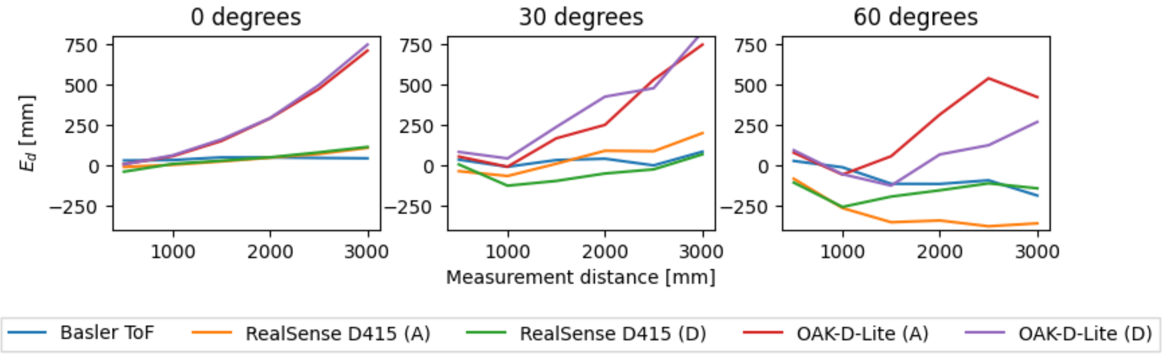


Figure 5.6 Distance estimation error measured on a concrete wall dataset

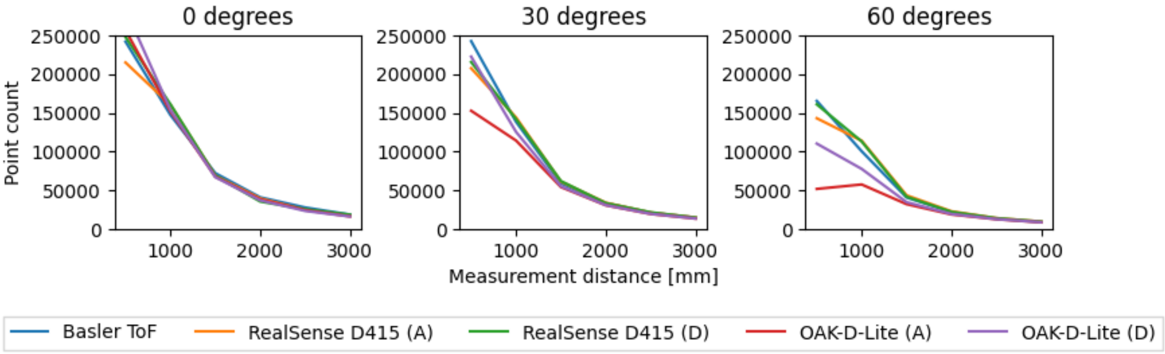


Figure 5.7 Mean amount of points captured on a concrete wall.

5.4 Concave surface profile measurement

The OAK-D-Lite camera was not able to discern the concave region from the top face due to a lack of distinct key points, as the measured objects all have a uniform black colour.

Figure 5.8 shows the profile errors measured on all of the cuboidal artefacts.

Both the Basler ToF and RealSense D415 were able to estimate the depth of the concave region reasonably well. The influence of measurement distance in the chosen range is substantial only on the artefact with dimensions of the concave region of 10 mm by 10 mm.

5.4.1 Influence of the width of the concave region

Figure 5.9 shows the influence of the width of the concave region on the profile estimation error. The mean error E_p is marked with a cross.

The Basler ToF camera could discern narrow concave regions with reasonable accuracy. Due to the density of its projected pattern, the RealSense D415 could not scan very narrow cuts. However, with the increasing width of the concave region, both of the cameras showed similar accuracy.

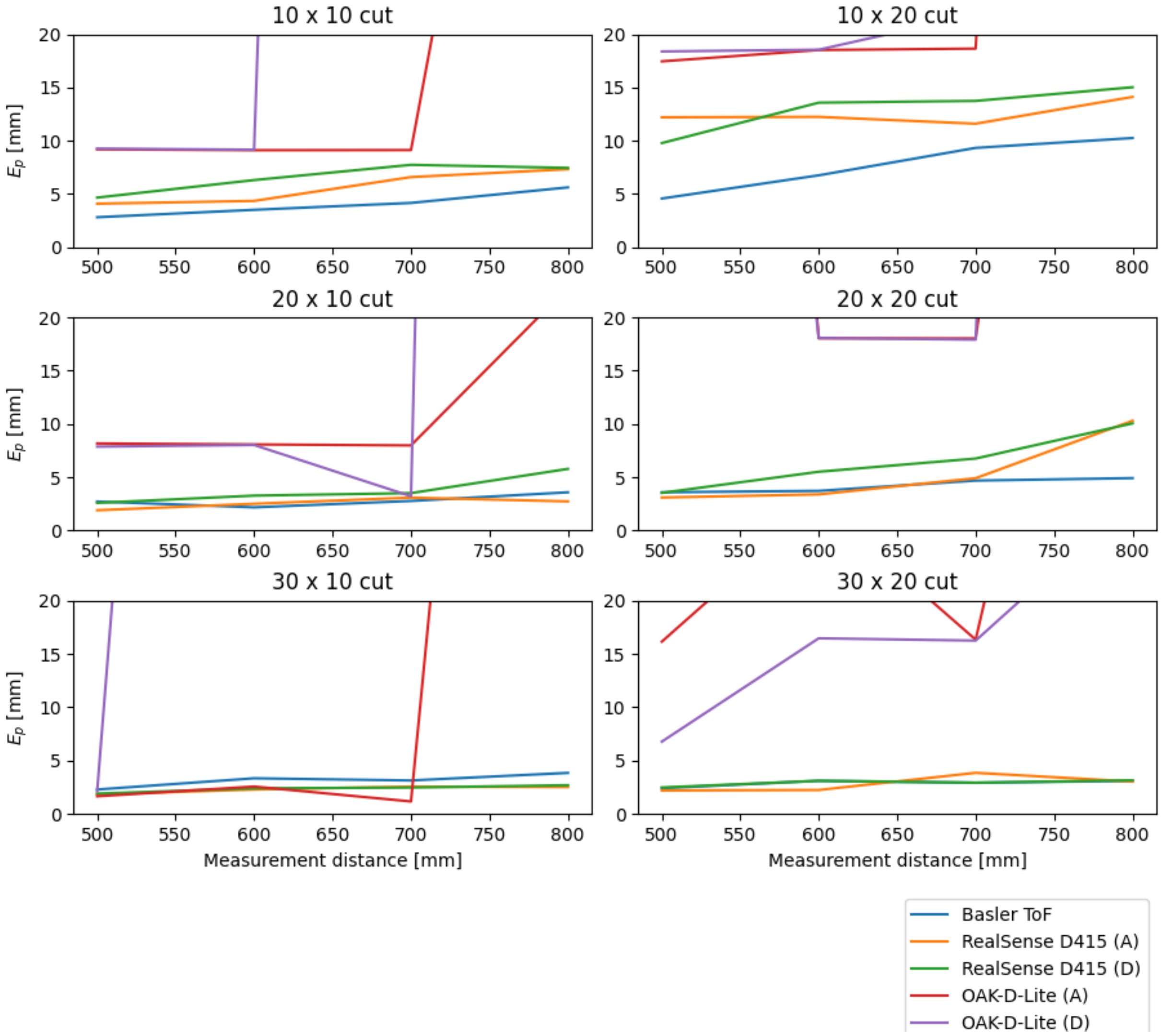


Figure 5.8 Profile error measured on all cuboidal artefacts.

5.5 Free-form artefact reconstruction

As mentioned in section 4.7, only the RealSense D415 camera was able to reconstruct the free-form artefact. Figure 4.21 shows the reconstructed artefact, figure 5.10 shows the histogram of point errors.

The scans from the camera with High Accuracy and High Density presets contain 4993 and 5561 points, respectively. The mean error E_m is 2.0 mm for both.

It is apparent from the error histogram that the presets did not influence the overall result. The lack of preset influence may be partly caused by the implementation of the point cloud fusion pipeline. The difference between the presets may also be better measured on a larger and a more complicated free-form artefact.

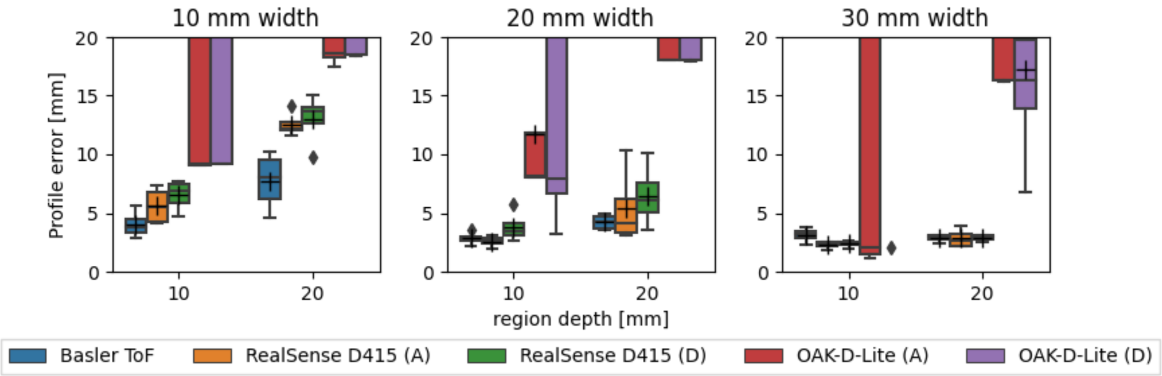


Figure 5.9 Influence of the concave region width on the measurement error.

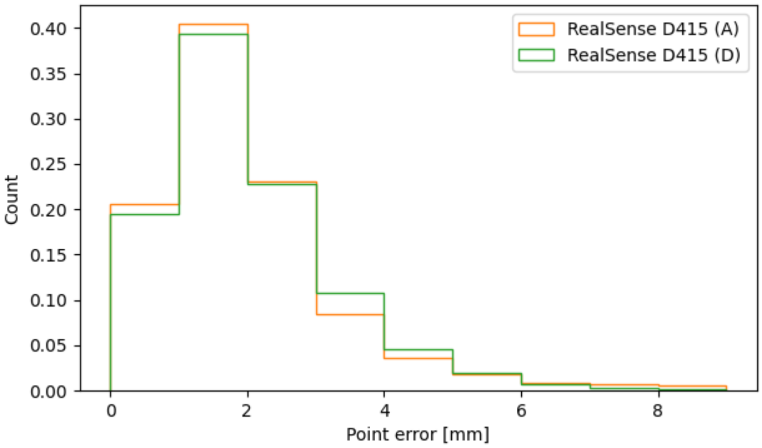


Figure 5.10 Reconstructed free-form artefact point error histogram

5.6 Measurement conclusion

Considering the measured errors, the following decision diagrams for selecting a depth-sensing device for computer vision tasks are presented. Figures 5.11 and 5.12 show diagrams for selecting a device for distance measurement and shape measurement, respectively.

Other factors such as operating conditions and environment, required FPS, FOV, scene lighting, host device, and processing pipeline requirements, to name a few, should still be considered carefully.

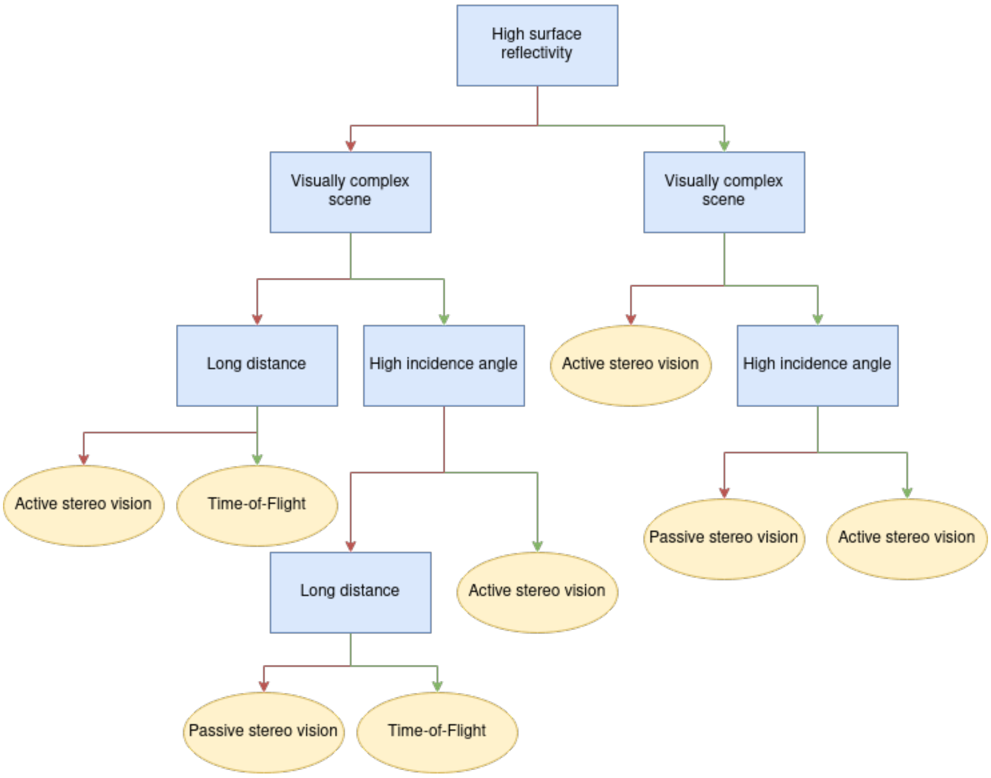


Figure 5.11 Decision diagram for selecting a depth-sensing device for distance measurement.

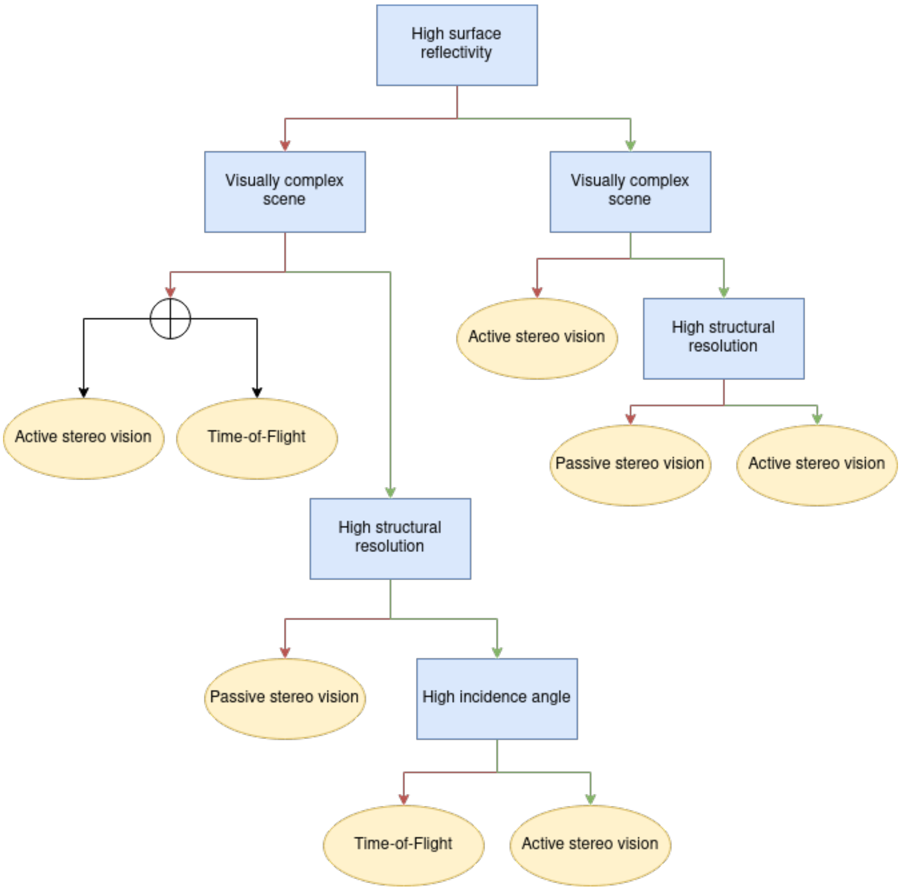


Figure 5.12 Decision diagram for selecting a depth-sensing device for shape measurement.

Chapter 6 Discussion

The Basler ToF camera performed reasonably well in all of the measured tests, although it is clearly not well suited for imaging reflective surfaces that cause a specular reflection. Continuation of the distance estimation tests seems interesting, as the reason for overestimating the wooden board distance when measuring under a high incidence angle is not clear. The fact that a similar systematic error was not present in the exposed concrete data hints at the surface characteristics of wood. It is possible that the coarse surface caused a diffusing reflection. The limitations of Time-of-Flight sensors seem very severe, and their extent should be investigated further.

The Intel RealSense D415 was quite accurate in all of the measurements. Neither of the tested influencing factors had any significant effect on its accuracy. Therefore, active stereo vision seems to be a good choice for accurate depth sensing in the low to medium range with high structural resolution regardless of the imaged objects.

The OAK-D-Lite was not well suited for most carried out tests due to a lack of a dense array of visually distinct key points present in the images. However, it performed reasonably well when imaging an exposed concrete wall, which clearly met the required conditions for dense disparity matching. Therefore it is expected that the camera would perform satisfactorily in well lit and visually complex environments, such as outside, to estimate the depth of scene objects. It is also quite exciting compared to the other cameras, as it can run image processing workloads on device in real-time. If such functionality is required, however, the depth accuracy and structural resolution are not satisfactory, Luxonis also offers an active version.

Chapter 7
Conclusion

Several depth camera comparison tests and their methodologies were created. Each of the tests measured the influence of measurement factors on the accuracy of the devices. The tests monitored such factors as the incidence angle, increasing distance, object surface characteristics, or object dimensions to establish the respective strengths, weaknesses, metrological characteristics, and the compared devices' structural resolutions. Three selected devices utilising different principles for estimating the scene depth were compared in each test. The principles were Time-of-Flight and active and passive stereo vision. After analysing the test results, decision diagrams and recommendations for the selection of a depth-sensing device were presented as a basis for future feasibility studies of computer vision projects utilising depth-sensing cameras.

Bibliography

1. SIMEK, Kyle. Dissecting the Camera Matrix, Part 3: The Intrinsic Matrix. In: *Dissecting the Camera Matrix* [online]. 2013 [visited on 2022-04-13]. Available from: `https://ksimek.github.io/2013/08/13/intrinsic/`.
2. ZHANG, Zhengyou. *International Journal of Computer Vision*. 1998, vol. 27, no. 2, pp. 161–195. Available from DOI: `10.1023/a:1007941100561`.
3. OPENCV CONTRIBUTORS. Epipolar Geometry. In: *Camera Calibration and 3D Reconstruction* [online]. 2015 [visited on 2022-04-22]. Available from: `https://docs.opencv.org/4.x/da/de9/tutorial_py_epipolar_geometry.html`.
4. FUSIELLO, Andrea; TRUCCO, Emanuele; VERRI, Alessandro. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*. 2000, vol. 12, no. 1, pp. 16–22. Available from DOI: `10.1007/s001380050120`.
5. OPENCV CONTRIBUTORS. Depth Map from Stereo Images. In: *Camera Calibration and 3D Reconstruction* [online]. 2015 [visited on 2022-04-22]. Available from: `https://docs.opencv.org/4.x/dd/d53/tutorial_py_depthmap.html`.
6. INTEL. Post processing filters. In: *Intel RealSense Documentation* [online]. 2019 [visited on 2022-04-27]. Available from: `https://dev.intelrealsense.com/docs/post-processing-filters`.
7. PCL CONTRIBUTORS. The PCD (Point Cloud Data) file format. In: *Point Cloud Library Documentation* [online]. 2018 [visited on 2022-04-22]. Available from: `https://pcl.readthedocs.io/projects/tutorials/en/latest/pcd_file_format.html`.
8. OPEN3D CONTRIBUTORS. open3d::geometry::PointCloud Class Reference. In: *Open3D Documentation* [online]. 2022 [visited on 2022-04-23]. Available from: `http://www.open3d.org/docs/release/cpp_api/classopen3d_1_1geometry_1_1_point_cloud.html`.
9. BENTLEY, Jon Louis. Multidimensional binary search trees used for associative searching. *Communications of the ACM*. 1975, vol. 18, no. 9, pp. 509–517. Available from DOI: `10.1145/361002.361007`.
10. BESL, P.J.; MCKAY, Neil D. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1992, vol. 14, no. 2, pp. 239–256. Available from DOI: `10.1109/34.121791`.
11. DECKER, Nathan; WANG, Yuanxiang; HUANG, Qiang. Efficiently registering scan point clouds of 3D printed parts for shape accuracy assessment and modeling. *Journal of Manufacturing Systems*. 2020, vol. 56, pp. 587–597. Available from DOI: `10.1016/j.jmsy.2020.04.001`.

12. FISCHLER, Martin A.; BOLLES, Robert C. Random sample consensus. *Communications of the ACM*. 1981, vol. 24, no. 6, pp. 381–395. Available from DOI: `10.1145/358669.358692`.

13. GENG, Jason. Structured-light 3D surface imaging: a tutorial. *Advances in Optics and Photonics*. 2011, vol. 3, no. 2, p. 128. Available from DOI: `10.1364/aop.3.000128`.

14. SAYYAR-ROUDSARI, Sajjad; HAMOUSH, Sameer A.; SZETO, Taylor M. V.; YI, Sun. Using a 3D Computer Vision System for Inspection of Reinforced Concrete Structures. In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, 2019, pp. 608–618. Available from DOI: `10.1007/978-3-030-17798-0_49`.

15. JOVANČEVIĆ, Igor; PHAM, Huy-Hieu; ORTEU, Jean-José; GILBLAS, Rémi; HARVENT, Jacques; MAURICE, Xavier; BRÈTHES, Ludovic. 3D Point Cloud Analysis for Detection and Characterization of Defects on Airplane Exterior Surface. *Journal of Nondestructive Evaluation*. 2017, vol. 36, no. 4. Available from DOI: `10.1007/s10921-017-0453-1`.

16. MAKUCH, Maria; GAWRONEK, Pelagia. 3D Point Cloud Analysis for Damage Detection on Hyperboloid Cooling Tower Shells. *Remote Sensing*. 2020, vol. 12, no. 10, p. 1542. Available from DOI: `10.3390/rs12101542`.

17. BAEK, Dongyoub; CHO, Sungmin; BANG, Hyunwoo. Wheel alignment inspection by 3D point cloud monitoring. *Journal of Mechanical Science and Technology*. 2014, vol. 28, no. 4, pp. 1465–1471. Available from DOI: `10.1007/s12206-014-0133-3`.

18. ZHOU, Song; XIAO, Sheng. 3D face recognition: a survey. *Human-centric Computing and Information Sciences*. 2018, vol. 8, no. 1. Available from DOI: `10.1186/s13673-018-0157-2`.

19. ZENGELER, Nico; KOPINSKI, Thomas; HANDMANN, Uwe. Hand Gesture Recognition in Automotive Human–Machine Interaction Using Depth Cameras. *Sensors*. 2018, vol. 19, no. 1, p. 59. Available from DOI: `10.3390/s19010059`.

20. BISWAS, Joydeep; VELOSO, Manuela. Depth camera based indoor mobile robot localization and navigation. In: *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012. Available from DOI: `10.1109/icra.2012.6224766`.

21. SOUZA, Joao Pedro C. de; ROCHA, Luis F.; FILIPE, Vitor M.; BOAVENTURA-CUNHA, J.; MOREIRA, A. Paulo. Low-Cost and Reduced-Size 3D-Cameras Metrological Evaluation Applied to Industrial Robotic Welding Operations. In: *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2021. Available from DOI: `10.1109/icarsc52212.2021.9429788`.

22. DEPTHAI CONTRIBUTORS. StereoDepth. In: *DepthAI Documentation* [online]. 2022 [visited on 2022-04-14]. Available from: `https://docs.luxonis.com/projects/api/en/latest/components/nodes/stereo_depth/#min-stereo-depth-dista`.

23. DEPTHAI CONTRIBUTORS. Depth Perception. In: *DepthAI Documentation* [online]. 2022 [visited on 2022-04-14]. Available from: `https://docs.luxonis.com/en/latest/pages/depth/#depth-perception`.

24. LI, Larry. *Time-of-Flight Camera - An Introduction* [online]. 2014 [visited on 2022-04-22]. No. SLOA190B. Available from: `https://www.ti.com/product/OPT8320`. Rev. B.

25. INTEL. Depth Camera D415. In: *Intel RealSense* [online]. 2018 [visited on 2022-04-27]. Available from: `https://www.intelrealsense.com/depth-camera-d415/`.

26. BASLER. Basler's First 3D Camera Enters Series Production. In: *Basler blog* [online]. 2016 [visited on 2022-04-20]. Available from: `https://www.baslerweb.com/en/company/news-press/press-releases/baslers-first-3d-camera-enters-series-production/12032/`.

27. DEPTHAI CONTRIBUTORS. OAK-D-Lite. In: *DepthAI Documentation* [online]. 2021 [visited on 2022-04-27]. Available from: `https://docs.luxonis.com/projects/hardware/en/latest/pages/DM9095.html`.

28. STANDARDIZATION, International Organization for. *Geometrical product specifications (GPS) — Acceptance and reverification tests for coordinate measuring systems (CMS) — Part 13: Optical 3D CMS*. ISO 10360-13:2021. Vernier, Geneva, Switzerland: International Organization for Standardization, 2021. Available also from: `https://www.iso.org/standard/74957.html`.

29. BASLER. The New Basler ToF Software Release Has Arrived. In: *Basler blog* [online]. 2018 [visited on 2022-04-20]. Available from: `https://www.baslerweb.com/en/company/news-press/news/the-new-basler-tof-software-release-has-arrived/39040/`.

Contents of the attached media

- `code`
  - `analysis` Jupyter notebooks for analysis of the measured data
  - `cam_grabbers` RealSense and OAK acquisition scripts
  - `preprocessing` format conversion, preprocessing, and segmentation scripts and utilities
  - `visualization` Jupyter notebooks for visualizations
- `data` directory containing the measured and reference data
- `thesis` directory containing the source files of this document
  - `thesis.pdf` this document