Estimating Head Measurements from 3D Point Clouds

Estimating Head Measurements from 3D Point Clouds

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M. Sc. Isabel Cristina Patiño Mejía

aus Cali, Kolumbien

Tübingen

2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	18.10.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. rer. nat. Andreas Zell
2. Berichterstatter:	Prof. Dr.-Ing. Hendrik P. A. Lensch

To my parents and my sister and to all the people that have crossed my path, without you I would not be where I am.

To HT and PT, thanks for all the emotional support.

Para mis padres y mi hermana y todas las personas que han cruzado mi camino, sin ustedes no estaría donde estoy.

Abstract

Human head measurements are valuable in ergonomics, acoustics, medicine, computer vision, and computer graphics, among other fields. Such measurements are usually obtained using entirely or partially manual tasks, which is a cumbersome practice since the level of accuracy depends on the expertise of the person that takes the measurements. Moreover, manually acquired measurements contain less information from which new measurements can be deduced when the subject is no longer accessible. Therefore, in order to overcome these disadvantages, an approach to automatically estimate measurements from 3D point clouds, which are long-term representations of humans, has been developed and is described in the presented manuscript. The 3D point clouds were acquired using an RGBD sensor Asus Xtion Pro Live and KinFu (open-source implementation of KinectFusion). Qualitative and quantitative evaluations of the estimated measurements are presented. Furthermore, the feasibility of the developed approach was evaluated through a case study in which the estimated measurements were used to appraise the influence of anthropometric data on the computation of the interaural time difference.

Considering the promising results obtained from the estimation of measurements from 3D models acquired with the sensor Asus Xtion Pro Live and KinFu (plus the results reported in the literature) and the development of new RGBD sensors, a study of the influence of seven different RGBD sensors on the reconstruction obtained with KinFu is also presented. This study contains qualitative and quantitative evaluations of reconstructions of four diverse objects captured at different distances that range from 40 cm to 120 cm. Such range was established according to the operational range of the sensors. Furthermore, a collection of obtained reconstructions is available as a dataset in http://uni-tuebingen.de/en/138898.

Kurzfassung

Maße menschlicher Köpfe sind unter anderem nützlich für die Ergonomie, die Akustik, die Medizin, Computer Vision sowie Computergrafik. Solche Maße werden üblicherweise gänzlich oder teilweise manuell gewonnen, was ein umständliches Verfahren darstellt, da die Genauigkeit von der Kompetenz der Person abhängt, die diese Messungen vornimmt. Darüber hinaus enthalten manuell erfasste Daten weniger Informationen, von denen neue Maße abgeleitet werden können, wenn das Subjekt nicht länger verfügbar ist. Um diese Nachteile wettzumachen, wurde ein Verfahren entwickelt, das in diesem Manuskript vorgestellt wird, um automatisch Maße aus 3D Punktwolken zu bestimmen, da diese eine langfristige Repräsentation von Menschen darstellen. Diese 3D Punktwolken wurden mit dem ASUS Xtion Pro Live RGB-D Sensor und KinFu (der open-source Implementierung von KinectFusion) aufgenommen. Es werden sowohl qualitative als auch quantitative Auswertungen der gewonnenen Maße präsentiert. Weiterhin wurde die Umsetzbarkeit des entwickelten Verfahrens anhand einer Fallstudie beurteilt, in der die gewonnenen Maße genutzt wurden, um den Einfluss von anthropometrischen Daten auf die Berechung der interauralen Zeitdifferenz zu schätzen.

In Anbetracht der vielversprechenden Ergebnisse der Bestimmung von Maßen aus 3D Modellen, die mit dem Asus Xtion Pro Live Sensor und KinFu erstellt wurden, (sowie der Ergebnisse aus der Literatur) und der Entwicklung neuer RGB-D Sensoren, wird außerdem eine Studie des Einflusses von sieben verschiedenen RGB-D Sensoren auf die Rekonstruktion mittels KinFu dargestellt. Diese Studie enthält qualitative und quantitative Auswertungen von Rekonstruktionen vier verschiedener Objekte, die in unterschiedlichen Distanzen von 40 cm bis 120 cm aufgenommen wurden. Diese Spanne wurde anhand der Reichweite der Sensoren gewählt. Des Weiteren ist eine Sammlung der erhaltenen Rekonstruktionen als Datensatz verfügbar unter http://uni-tuebingen.de/en/138898.

Acknowledgments

I thank Prof. Zell for the opportunity to obtain my doctoral degree in his chair, for the support and the interesting discussions that contributed to the elaboration of this thesis. I thank all my colleagues, the ones that are still here and the ones that are already gone. All of you collaborated directly or indirectly with my work. Additionally, I show my gratitude to all the people that participated in the experiments.

I acknowledge the support of the German Federal Ministry of Education and Research (BMBF), who funded my initial project. I thank also the partners of such project: Christian Hoene from Symonics, Ramona Bomhardt and Prof. Fels from the RWTH Aachen.

Finally, I appreciate all the help and support that I received from Thiago, from my parents, my sister and my family in general, as well as from my friends; without you, this path would have been harder.

Contents

Chapter 1

Introduction

1.1 Motivation and Contribution

Head measurements are useful in many fields, such as ergonomics, medicine, computer vision, computer graphics, and acoustics, among others. The specific required measurements depend on their availability and application. The manual acquisition of anthropometric dimensions takes time, and its accuracy depends on the expertise of the person that measures. Moreover, the measurement acquisition experiments need to be very well planned, since the reduced available information diminishes the possibility of obtaining new post-hoc measurements. In order to overcome this obstacle, and taking advantage of the improvement of science and technology, approaches to acquire anthropometric dimensions from long-term representations of humans, like images [97] and 3D models [35, 76] have emerged; such 3D models are usually represented as meshes and point clouds. The presented manuscript focuses on the estimation of head measurements from 3D point clouds, which are structures, where each point has three coordinates (x, y, z) that locate it in space and can additionally contain color information.

The aforementioned point clouds are estimated using information captured with either monocular cameras or depth sensors, such as stereo vision systems, time-of flight and structured-light based sensors. This information is usually processed through approaches based on the detection and tracking of salient points (features such as edges and corners) in order to obtain a 3D representation of the world. However, methods based on information fusion, without salient points extraction, are nowadays also an accurate alternative. This kind of approaches started with Microsoft KinectFusion [70, 48], which was developed to work with an RGBD (red, green, blue, depth) sensor called Microsoft Kinect. This sensor is based on structured-light technology to recover depth information of the scene, which together with the color information captured by the RGB camera offers a complete spatial and color representation of the world. The Microsoft Kinect was a pioneer on this kind of out-of-the-shelf sensors; after it was released, multiple other devices have been developed with similar technology, such as the Xtion Pro and Pro Live created by Asus, and the Astra and Astra S developed by Orbbec. Due to the use of projected patterns, a disadvantage of the structured-light technology is that its use is limited to indoor environments. Therefore, Intel has been developing sensors that use stereo vision

systems with a laser projector to enhance depth acquisition, such as the RealSense R200 and D415. Since the depth acquisition algorithm of such sensors does not depend on the projected laser, they are suitable to work in outdoor environments.

Considering the great performance of the RGBD sensors and the promising results given by fusion algorithms like KinectFusion, a combination of these two technologies assures a valuable alternative to classical approaches for the 3D reconstruction of the world. Therefore, this is the combination employed to generate the 3D point clouds used on the automatic estimation of head measurements. For this purpose, a state-of-the-art (at this time) RGBD sensor called Asus Xtion Pro Live and the software KinFu [4] (open-source implementation of KinectFusion available in the point cloud library - PCL - [8]) were used. The developed algorithm is based on three point clouds of the upper part of the human body: a 360° reconstruction and two profile point clouds (right and left). The measurements to be estimated were determined by the requirements to generate personalized Head Related Transfer Functions [19] (a person-dependent function used to localize sound sources - HRTF -). Such measurements are the head width, depth, and height, as well as the closest point from the front and back of the head to the ear pits. Experiments with 20 subjects were performed in order to evaluate the developed algorithm. Additionally, the feasibility of the approach was evaluated through a case study, where the estimated measurements were employed to assess the influence of anthropometric information on the interaural time difference [89].

The combination of an RGBD sensor with algorithms like KinFu proved to be a great alternative to generate 3D reconstructions to automatically extract anthropometric data. Moreover, the 3D reconstructions can also be used to create datasets such as BigBIRD [91], and A Large Dataset of Object Scans [31]. Nevertheless, nowadays so many RGBD sensors exist that it is significant to study if the generated 3D reconstructions depend on the sensor used to acquire the input data. Hence, a qualitative and quantitative study is presented in this manuscript. In the study, four objects were reconstructed using point clouds recorded with six well-known RGBD sensors and KinFu, this algorithm was preferred for its flexibility regarding sensors and offline usability. The sensors used were: a Microsoft Kinect Xbox 360, a Microsoft Kinect in near mode, an Asus Xtion Pro Live, an Orbbec Astra S, and an Intel RealSense R200. The reconstructed objects were a dummy head, a soccer ball, and American Football, and a large rubber duck. An additional preliminary study was performed on the reconstructions of the dummy head created from point clouds recorded with an Intel RealSense D415. Since the experiments involved in such studies resembled scenarios that researchers and end users might encounter, the results of these experiments represent a technically-grounded guideline for RGBD sensor selection. Furthermore, a collection of obtained reconstructions is available as a dataset in http://uni-tuebingen.de/en/138898.

1.2 Structure of the Manuscript

The remaining of this manuscript is organized as follows. Chapter 2 presents a description of a camera model and depth sensors, including stereo vision systems and out-of-the-shelf RGBD sensors. The description of approaches used to generate 3D reconstructions of the world as a rigid and non-rigid environment, with either features extraction or information fusion is presented in Chapter 3. The algorithm developed to automatically estimate head measurements from 3D points clouds, as well as its evaluation is presented in Chapter 4. This chapter is based on the publication

- Patino Mejia, I., and Zell, A. Head Measurements from 3D Point Clouds. *6th International Conference on Image Processing Theory, Tools and Applications (IPTA 2016)*. 2017 [76].

Part of the described algorithm contributed also to the co-authored publication

- Hoene, C., Patino Mejia, I., and Cacerovschi, A. MySofa—Design Your Personal HRTF. *142nd Audio Engineering Society Convention*. 2017 [47].

Additionally, the case study used to evaluate the feasibility of the algorithm that automatically estimates head measurements is described in Chapter 5. It is based on the shared publication

- Bomhardt, R., Patino Mejia, I., Zell, A., and Fels, J. Required Measurement Accuracy of Head Dimensions for Modeling the Interaural Time Difference[1]. *Journal of the Audio Engineering Society*. 2018 [26].

Furthermore, Chapter 6 is based on the publication

- Patino Mejia, I., and Zell, 3D Reconstructions with KinFu Using Different RGBD Sensors. *3rd IEEE International Conference on Image Processing, Applications and Systems (IPAS 2018)*. 2018 [77],

it contains the analysis performed to 3D reconstructions generated with different RGBD sensors. Conclusions and future work are found in Chapter 7.

[1] The first and second authors contributed equally to this work.

Chapter 2

Cameras and Sensors

Along the years the world has been captured in 2D and 3D using different devices and techniques. In this chapter, a brief explanation of the most classical camera model to represent 3D scenes in 2D is briefly described in Section 2.1. Furthermore, since this manuscript focuses on 3D representations of humans and objects, a description of sensors specialized in generating such representations is presented in Section 2.2.

2.1 Pinhole Camera Model

As explained in [46], a camera maps 3D information to 2D images, i.e., the dimension of the environment is reduced through a projection process. The most basic camera model known as the pinhole camera model maps a point in the 3D world with coordinates $\vec{X} = (X,Y,Z)^T$ to the image plane, drawing a line between the 3D point and the camera center (see Fig. 2.1).

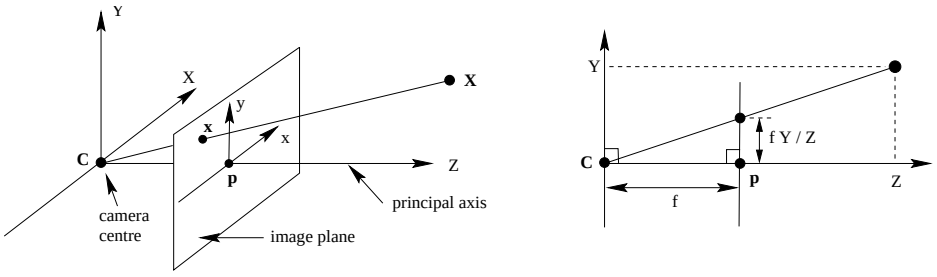


Figure 2.1: The projection of a 3D point $\vec{X}$ into the image plane with principal point $\vec{p}$ connecting a ray between $\vec{X}$ and the camera center $\vec{C}$. Source: [46].

Considering that the link between the camera center and the principal point is the focal length f, the projection of the 3D point onto the 2D image is described (using homogeneous coordinates) as

$$\begin{bmatrix} fX \\ fY \\ Z \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} . \tag{2.1}$$

Since the origin of the image plane is not always on the principal point, a principal point offset (p_x, p_y) is added to eq. 2.1. Furthermore, the camera rotation and translation with respect to the world coordinate system is also required. Therefore, the mapping from 3D to 2D is solved by the equation

$$\vec{x} = K[R|t]\vec{X} , \tag{2.2}$$

where

$$K = \begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.3}$$

is known as the intrinsic camera parameters and $[R|t]$ represents the relative camera pose to the world, also called extrinsic camera parameters. The 3D and 2D points are symbolized by $\vec{X}$ and $\vec{x}$, respectively.

2.2 Depth Sensors

In order to obtain a realistic reconstruction of the world, it is necessary to acquire depth information of it. Such information can be acquired directly using special sensors with either active or passive technology. Active methods directly interact with the scene to reconstruct, e.g., projecting lasers to obtain the depth data. On the other hand, passive methods do not directly interact with the scene and rely on image processing techniques to extract the depth, e.g., stereo vision systems.

The depth information obtained from the sensors can be represented as depth maps or range images, point clouds, and meshes. In this section, a brief explanation of active and passive sensors used to obtain depth information of the world is presented.

2.2.1 Stereo vision systems

Binocular vision is the human (and some other animals') characteristic that allows the perception of distances. Both eyes send information to the brain at the same time, the information given by one eye is slightly different than the one given by the other eye. Such differences are processed by the brain with the aim of indicating the distances to the surrounding elements [30]. This principle is used in a stereo vision system, in which two cameras are separated by a baseline simulating the human eyes as is shown in Fig.

2.2. The baseline length determines the depth acquisition range. With a short baseline, it is possible to acquire depth at shorter distances than with a system with a longer baseline [68]. In order to acquire the depth data, corresponding points in the stereo (left - right) images are required. The difference in the position of such corresponding points is called disparity [21]. Once the disparity map is generated, a 3D point cloud can be obtained. Examples of point clouds obtained with a stereo vision system Nerian sp1 are shown in Fig. 2.3. The scene to be reconstructed is a dummy head placed at different distances in a room illuminated with yellow incandescent ceiling lamps. The point clouds are noisy, and the wall in the back (located no more than 3 m away) is not densely reconstructed.

Figure 2.2: Stereo vision systems. Top: 10 cm baseline. Bottom: 25 cm baseline. Source: [68].

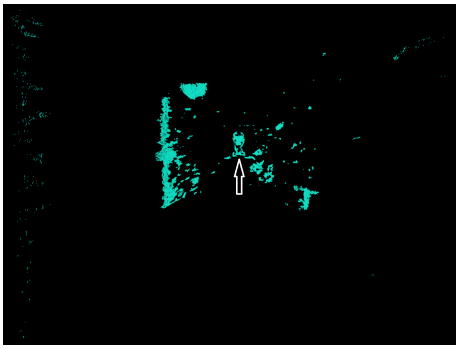
2.2.2 Out-of-the-shelf RGBD sensors

The presented manuscript is based on data acquisition with out-of-the-shelf RGBD sensors. The low price and ability of these sensors to capture the 3D world make them valuable in a wide range of applications, e.g., object and people detection and identification, motion recognition, and 3D reconstruction. In this section, a brief description of such sensors is presented.

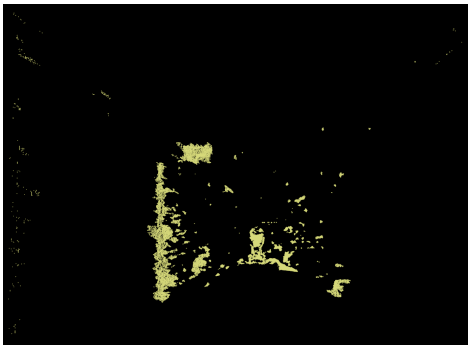
Structured-light sensors

Structured light is a technique based on the extraction of depth measurements from projected-light patterns like the systems used in [44, 87] to reconstruct surfaces using a setup of cameras and projectors. However, those configurations contain the cameras and the projectors as two separate entities. The company PrimeSense together with Microsoft popularized such technology by embedding it in a single device: the Microsoft Kinect, which was launched in 2010 as a peripheral for the videogame console Microsoft Xbox 360. The application of the Kinect in research projects motivated the development of more sensors with the same technology, like the Microsoft Kinect in near mode, the

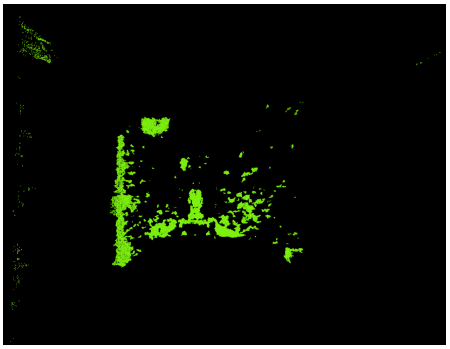
(a) Dummy head.



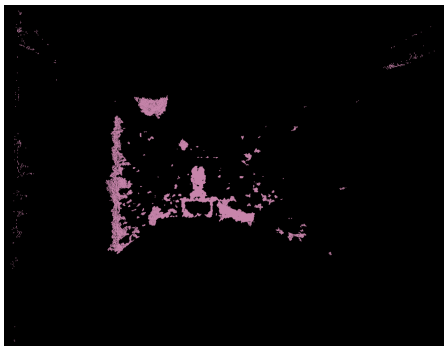
(b) 50 cm.



(c) 60 cm.



(d) 80 cm.



(e) 100 cm.

Figure 2.3: Examples of reconstructions of a dummy head with a stereo vision system. The dummy head is placed at different distances and noisy point clouds are recorded. The arrow gives a first guide to the reader to the reconstruction of the dummy head, it does not belong to the point cloud. The point clouds' colors are just for rendering purpose.

PrimeSense Carmine, the Asus Xtion Pro Live, and the Orbbec Astra, among others. Intel has also ventured in this field with sensors such as the RealSense F200, which uses a similar technology to the Kinect.

These sensors have an infrared (IR) projector, an infrared sensor, and an RGB camera

(see Fig. 2.4). In order to obtain depth measurements, the infrared projector emits a fixed pattern of spots that is projected onto the scene. The infrared sensor captures the image with the projected pattern. Then, the system compares the returned pattern with a reference, which has been recorded at a known distance. The differences between the patterns in the x axis permit the computation of the z (depth) values [39].

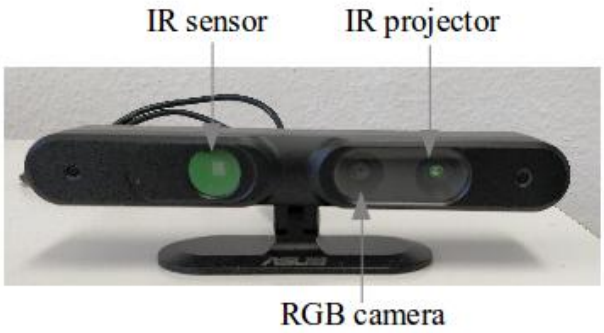


Figure 2.4: RGBD sensor using structured-light technology: Asus Xtion Pro Live.

In the presented work, the following sensors with structured-light technology were used: the Microsoft Kinect Xbox 360, the Microsoft Kinect in near mode (from now on called Kinect near mode), the Asus Xtion Pro Live, and the Orbbec Astra S. According to the manufacturers, the sensors Astra S acquires depth data from 40 cm to 200 cm [7] and the Kinect near mode from 40 cm to 300 cm [6]. The Xtion Pro Live and the Kinect Xbox 360 have a larger working distance range, from 80 cm to 350 cm [15] and 120 cm to 350 cm [11], respectively. A summary of the sensors' specifications (as given by the manufacturers) is shown in Table 2.1 and Table 2.2. The aforementioned sensors and their outcome can be managed through different alternatives, which include their own software development kits and open-source libraries. Point clouds generated using the Asus Xtion Pro Live and the ROS [13] package openni_launch [66, 65] are depicted in Fig. 2.5. The scene to reconstructs is the same as in Section 2.2.1: a dummy head (see Fig. 2.3a) is placed at different distances in a room illuminated with yellow incandescent ceiling lamps. The point clouds do not present significant noise, and the wall in the back (located no more than 3 m away) is densely reconstructed.

Time-of-flight sensors

Time of Flight (ToF) is a method for measuring the sensor-object distance using the time difference between the emission of a signal and its return to the sensor after its reflection on the object. The most commonly used signals are sound and light [14, 57].

The field of out-the-shelf inexpensive sensors was dominated by structured-light technology until 2014, when Microsoft presented the Kinect v2 as an evolution of the previously launched Kinect. To acquire depth information, the Microsoft Kinect v2 uses a time-of-flight camera that computes the distance to an object measuring the time of the

Specification	MS Kinect Xbox 360	MS Kinect Near Mode
Microphones	Array	Array
Depth image size	320×240 at 30 fps	320×240
Color image size	640×480 at 30 fps	640×480
Field of view	57° horizontal 43° vertical	57° horizontal 43° vertical
Distance of use	120 cm to 350 cm	40 cm to 300 cm
Operating system		Windows 7/8
Dimensions	27.9×7.1×6.6 cm	27.9×7.1×6.6 cm
Others	Motorized tilt Extra power supply needed	Motorized tilt Extra power supply needed

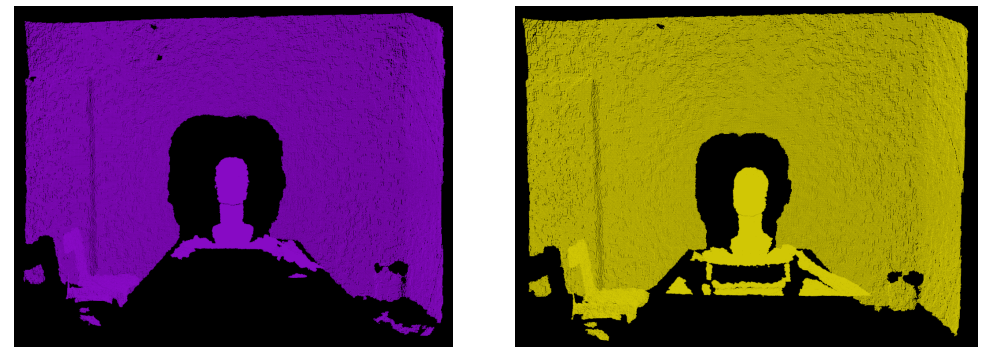
Table 2.1: Specifications of the Microsoft Kinect Xbox 360 and Kinect in near mode. Sources: [11, 6, 5, 95, 23, 92].

Specification	Asus Xtion Pro Live	Orbbec Astra S
Microphones	2	2
Depth image size	VGA (640×480) at 30 fps QVGA (320×240) at 60 fps	SXGA (1280×1024) at 5 fps* VGA (640×480) at 30 fps QVGA (320×240) at 30 fps QQVGA (160×120) at 30 fps
Color image size	SXGA (1280×1024)	1280×960 at 7 fps 640×480 at 30 fps 320×240 at 30 fps
Field of view	58° horizontal 45° vertical 70° diagonal	60° horizontal 49.5° vertical 73° diagonal
Distance of use	80 cm to 350 cm	40 cm to 200 cm
Interfaces	USB 2.0/ 3.0	USB 2.0
Operating system	Windows 32/64: XP, Vista, 7, 8 Linux Ubuntu 10.10: X86, 32/64 bit Android (by request)	Windows 7/8/10 Linux Android
Software	OpenNI	OpenNI 2
Dimensions	18×3.5×5 cm	16.5×3×4 cm

* Windows only.

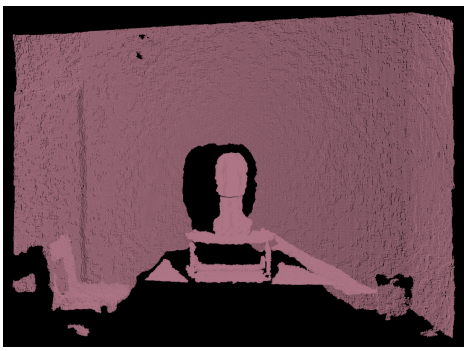
Table 2.2: Specifications of the Asus Xtion Pro Live and the Orbbec Astra S. Sources: [15, 7].

round trip travel of an amplitude-modulated light from the source to the object and back to the camera at each pixel. This kind of technology reaches accurate high pixel resolu-



(a) 60 cm.

(b) 80 cm.



(c) 100 cm.

(d) 120 cm.

Figure 2.5: Reconstruction of a dummy head with an Asus Xtion Pro Live. The dummy head is placed at different distances. The point clouds' colors are just for rendering purpose.

tion, and low motion blur, among others [78]. The Kinect v2 has an infrared source, an infrared sensor, and an RGB camera (see Fig. 2.6). Its depth range varies from 50 cm to 450 cm [95]. A summary of the sensor's specifications is shown in Table 2.3. The outcome of the Kinect v2 can be obtained through different alternatives, including its own software development kit and open-source libraries. Therefore, point clouds generated using this sensor and the ROS package IAI Kinect2 [102] are depicted in Fig. 2.7. The scene to be reconstructed is the same as in Section 2.2.1 and 2.2.2: a dummy head (shown in Fig. 2.3a) is placed at different distances in a room illuminated with yellow incandescent ceiling lamps. The point clouds, although dense and detailed, exhibit noise.

Stereoscopic sensors

Intel has developed sensors, such as the RealSense R200, D415 and D435, which are stereo vision systems enhanced with a pattern projector to improve the depth data acquisition. The projected pattern creates texture on the surface, generating unambiguous image matching. Due to their stereoscopic technology, these sensors are suitable to work

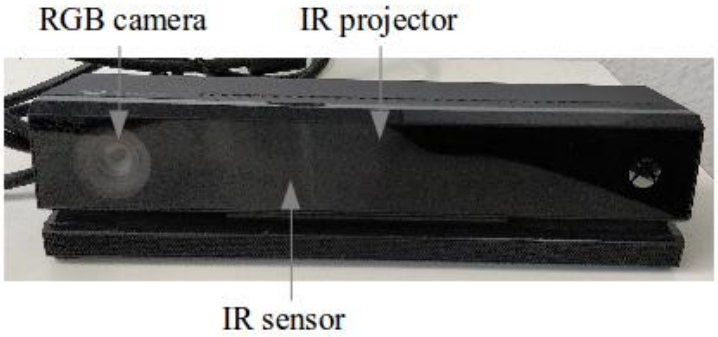
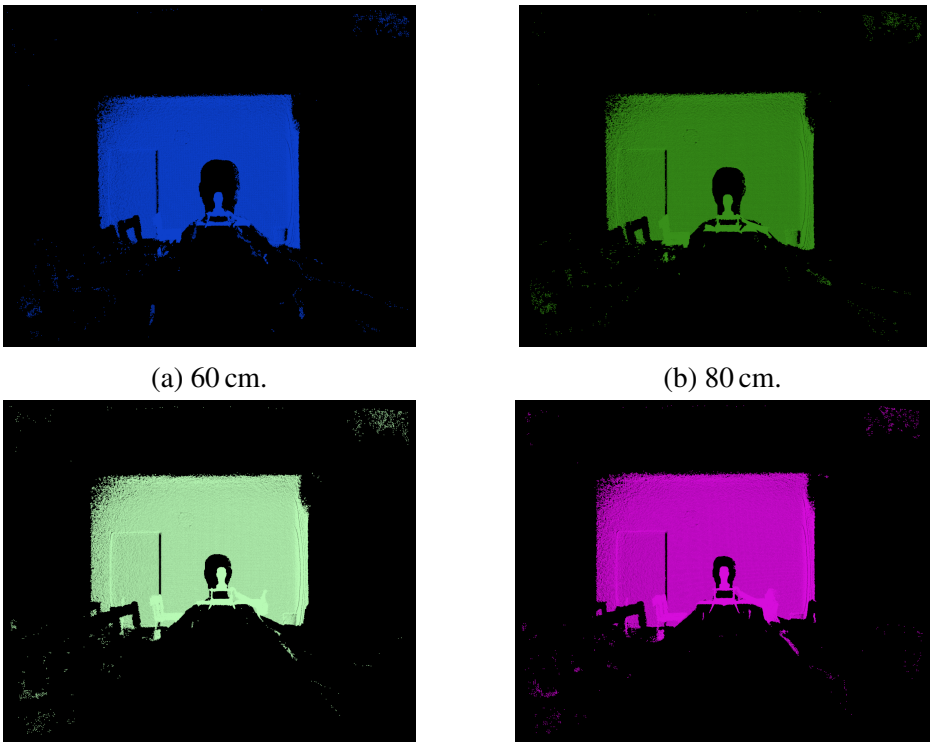


Figure 2.6: RGBD sensor using time-of-flight technology: Microsoft Kinect v2.

Specification	MS Kinect v2
Microphones	Array
Depth image size	512×424
Color image size	1920×1080 at 30 fps
Field of view	70° horizontal 60° vertical
Distance of use	50 cm to 450 cm
Interfaces	USB 3.0
Operating system	Windows 8
Dimensions	24.9×6.6×6.7 cm
Others	A hub and external power supply are required.

Table 2.3: Specifications of the Kinect v2. Sources: [95, 12, 93].

outdoors. In this work, the sensors R200 and the D415 were used. Both sensors have two cameras (right and left) to produce stereoscopic depth, an RGB camera, and an infrared projector to generate texture (see Fig. 2.8) [32, 51]. The depth range of the R200 varies from 50 cm to 350 cm indoors and up to 1000 cm outdoors [32], for the D415 the range is from 30 cm to 1000 cm [3]. Table 2.4 shows a summary of the R200's and D415's specifications (as given by the manufacturer). Additionally, point clouds captured with the R200 and the ROS package RealSense [9] are depicted in Fig. 2.9. The scene to be reconstructed is the same as in the sections above: a dummy head (depicted in Fig. 2.3a) is placed at different distances in a room illuminated with yellow incandescent ceiling lamps. The point clouds have some noise, and the wall in the back (located no more than 3 m away) is not densely reconstructed. Moreover, a decrement on the dummy head's quality acquisition is notable at 120 cm.



(a) 60 cm. (b) 80 cm.

(c) 100 cm. (d) 120 cm.

Figure 2.7: Reconstruction of a dummy head with a Microsoft Kinect v2. The dummy head is placed at different distances and noise is also recorded at each distance. The point clouds' colors are just for rendering purpose.

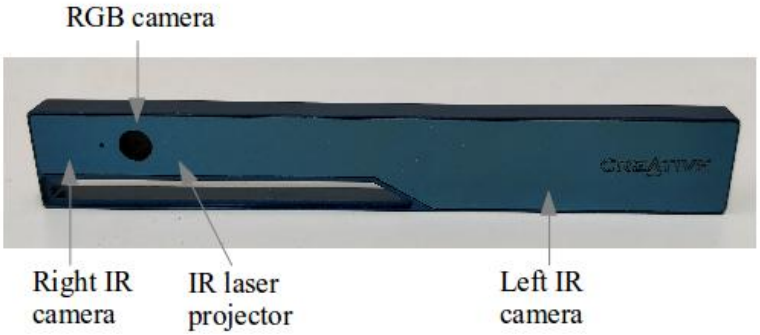
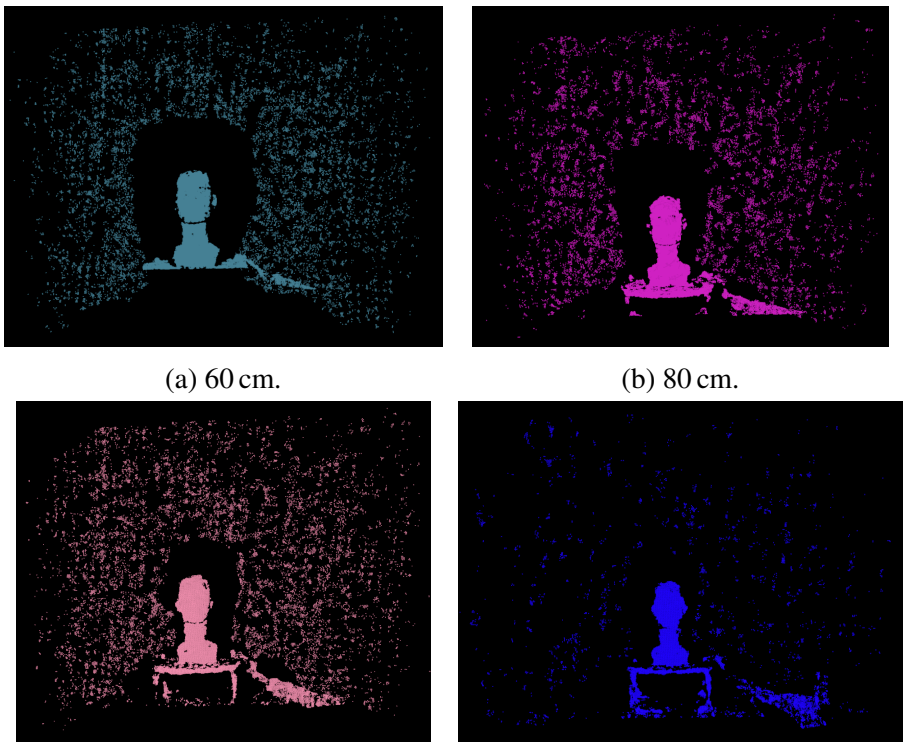


Figure 2.8: RGBD sensor with stereoscopic technology: Intel RealSense R200.

Specification	Intel RealSense R200	Inter RealSense D415
Right and left cameras image size	320×240 at 30 and 60 fps 480×360 at 30 and 60 fps	up to 1280×720 at up to 90 fps
Color image size	640×480 at 30 and 60 fps 1920×1080 at 30 fps	
Right and left cameras field of view	60° horizontal 45° vertical 70° diagonal	69.4° horizontal 42.5° vertical 77° diagonal
Color camera field of view	70° horizontal 43° vertical 77° diagonal	69.4° horizontal 42.5° vertical 77° diagonal
Distance of use	50 cm to 350 cm indoor 50 cm to 1000 cm outdoor	30 cm to 1000 cm
Interfaces	USB 3.0	USB 3.0
Dimensions	13×2×0.7 cm	9.9×2×2.3 cm

Table 2.4: Specifications of the sensors Intel RealSense R200 and D415. Sources: [32, 51, 3].



(a) 60 cm. (b) 80 cm.

(c) 100 cm. (d) 120 cm.

Figure 2.9: Reconstruction of a dummy head with an Intel RealSense R200. The dummy head is placed at different distances. The point clouds' colors are just for rendering purpose.

Chapter 3

3D Reconstruction Approaches

The 3D reconstruction of the world is useful in many fields, like anthropology, medicine, robotics, and computer vision, among others. Therefore, the development of methods to obtain such a reconstruction has been of interest to researchers since decades ago. In this chapter some of these approaches are presented. Section 3.1 describes methods related to structure from motion. Reconstructions of non-rigid scenes are presented in Section 3.2. Finally, methods that reconstruct scenes using depth sensors are described in Section 3.3.

3.1 Structure from Motion

Structure from motion (SfM) is the art of obtaining a 3D interpretation of the world based on 2D images. Such a world is assumed to have either rigid motion or a motion produced only by the camera who sees it. Along the years many approaches have been developed using both linear and nonlinear methods [49]. However, in general, this process involves three main stages [72]:

1. Feature extraction and matching: features are salient characteristics of the images, such as corners and edges. Several approaches for feature extraction exist, e.g., SIFT [60], SURF [22], ORB [88]. Each of these approaches presents qualities that make them suitable for different applications. However, it is important for a feature to be scale and rotation invariant.

2. Camera motion estimation: this is performed using the relative position between cameras estimated from the extracted features.

3. 3D reconstruction: using the extracted features and the estimated camera poses, the 3D world is reconstructed. In order to improve the accuracy, methods to minimize the reprojection error, like bundle adjustment [99], are used.

Structure from motion was used to reconstruct cities in [16]. The approach is based on an enormous quantity of images (taken from the internet), which generate unstructured photo collections. The photos in such collections present a challenge for SfM because

they are usually taken with different cameras, different illumination, and might not have any camera calibration information. The authors used a set of parallel distributed matching and reconstruction algorithms to reconstruct data sets of 150.000 images in less than a day (see Fig. 3.1). The system ran on a cluster of computers or nodes with a master node, which was responsible for the job scheduling. SIFT features were extracted and matched using the approximate nearest neighbor search (which uses k-d trees), the matches were verified using a RANSAC-based estimation of the fundamental or essential matrix. Usually, the feature matching is performed between two images. However, due to the number of images it was not feasible to compare every two pair of photos. Therefore, each image pair was selected using a multi-stage matching scheme based on vocabulary trees and query expansion. Then, a minimal subset of photos that covered the geometry of the scene was reconstructed, followed by the addition of the remaining photos using pose estimation, the triangulation of all the remaining points, and finally, the application of a customized bundle adjustment to refine the SfM.

(a) Dubrovnik.

(b) Rome.

Figure 3.1: Examples of cities' places reconstructed using SfM. Source: [16].

An approach to compute simultaneously extrinsic camera poses and 3D scene structure from high-resolution (2 megapixels to 20 megapixels) and high-frame-rate (25 Hz to 120 Hz) videos was presented in [82]. The system is also capable of integrating information from different videos. In order to obtain a globally consistent calibration and the 3D reconstruction of the scene, the system uses a modified 2D feature tracker (data coherence is used to reduce drift), followed by the application of a window bundle adjustment strategy on a set of confident frames. Additionally, global anchor links are established between different frame pairs and relative camera pose constraints are also generated.

Furthermore, at the end of the process, a global bundle adjustment is performed and all the less confident frames are added.

Other researches related to 3D reconstruction of rigid scenes can be found in [52, 45]. Moreover, a structure from motion method without correspondences is described in [34].

Simultaneous localization and mapping (SLAM) can be seen as a special case of SfM [72]. Approaches can be found in [67, 108, 109].

3.2 3D Representation of Non-Rigid Scenes

The real world is not only occupied by rigid objects, but there are several scenes where motion is presented, e.g., people talking, dancing, walking, animals moving. Those scenes are also valuable to reconstruct. However, classical methods as structure from motion fail on this task. Therefore, a different field has been developed in which a 3D reconstruction of the world is possible even if there is motion in it: 3D reconstruction of non-rigid objects.

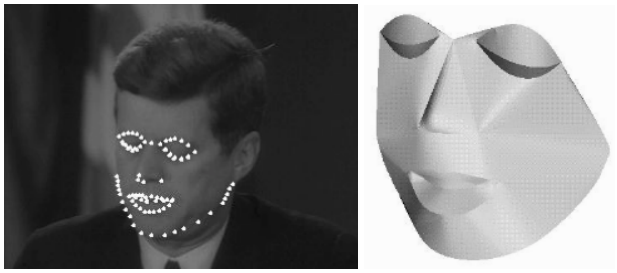
An approach to extract 3D non-rigid shape models from 2D image sequences was proposed in [27]. The algorithm works without an a-priori model and allowed, for example, the reconstruction of a 3D model of the human face, including facial expressions and lip movements from a video of the person talking. Additionally, 3D reconstructions of animals in motion were also obtained. The method is based on the representation of the 3D shape in each frame as a linear combination of basis shapes. The facial features are tracked using an appearance-based 2D tracking technique, animal features are tracked with a point feature tracker. Figure 3.2 shows some results of this approach.

A method to capture animals' articulated and deformable shape, as well as their texture from images and videos was proposed in [113]. The images are not assumed to be taken with either a still camera or the same camera, which increases the challenging level of the task. A strong prior model of the animal (a model of a toy figurine) is refined using information from images in which the animal portraits different poses. This approach uses the idea that separating the articulated structure of the animal from its shape, a consistent shape can be determined from the images. Such consistent shape is obtained using keypoints on the animal's body and silhouette and can be deformed again to match the respective poses on the images (see Fig. 3.3).

Other approaches to reconstruct non-rigid scenes can be found in [98, 40, 59].

3.3 3D Reconstruction with Depth Sensors

The 3D reconstruction of the world does not have to start always from 2D images. As explained in Chapter 2, sensors that give directly 3D information exist. In this section, some reconstruction methods based on such sensors are briefly explained.



(a) Reconstructed face (the points represent the tracked features).

(b) Reconstructed animal.

Figure 3.2: Examples of 3D reconstructions of non-rigid objects. Source: [27].

(a) Horse.

(b) Tiger.

Figure 3.3: 3D reconstruction of animals. Left: Real image; middle: 3D model; right: texturized model. Source: [113].

3.3.1 Stereo systems and laser-camera setups

An approach to densely reconstruct static scenes with high-resolution (up to 1 Megapixel) stereo sequences in real time was proposed in [42]. The process worked on two CPU cores and consists of: feature motion and egomotion estimation at 25 fps, dense stereo matching and 3D reconstruction at 3 fps to 4 fps. This algorithm requires a calibrated stereo system and rectified images as input. The feature matching is performed among four images (right and left images of two consecutive frames) at each time. The camera motion is obtained using a minimization of the sum of reprojection errors and a Kalman filter. The disparity maps are obtained using ELAS [41]. Outcome examples of this method are shown in Fig. 3.4.

(a) Right: some input images of a toy human. Left: 3D reconstruction.

(b) Three different viewpoints of a single sequence.

Figure 3.4: 3D reconstructions obtained with the method explained in [42]. The images were taken from the same publication.

A multi-view, multi-projector, and multi-pattern phase scanning method used for the 3D reconstruction of challenging materials and handling of occlusions was presented in [44]. The reconstruction method includes the projection and capture of shifted sinusoidal patterns (projected on the object to reconstruct) and bundle adjustment with generated features (needed for the system calibration). In order to determine the object surface, the structured light signal produced by each projector is captured by every camera, the place where all the structured light signals from different viewpoints align is considered as a point on the surface. Therefore, all the available cameras, projectors, and patterns

data are necessary to determine the depth of a surface point, this method is called multi-view optimization. All the calculations can be performed for each 3D point without information about the neighbors. Figure 3.5 shows the reconstruction of a wooden figure using the aforementioned approach.

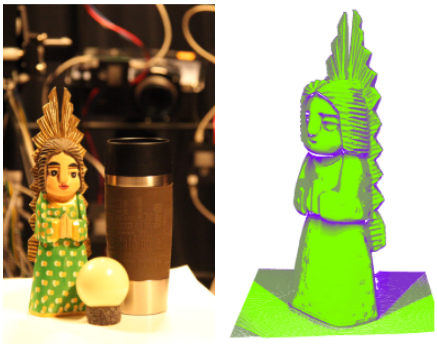


Figure 3.5: Reconstruction of a wooden figure. Left: Three different objects used during the experiments, the wooden figure is painted in green. Right: 3D reconstruction. Source: [44].

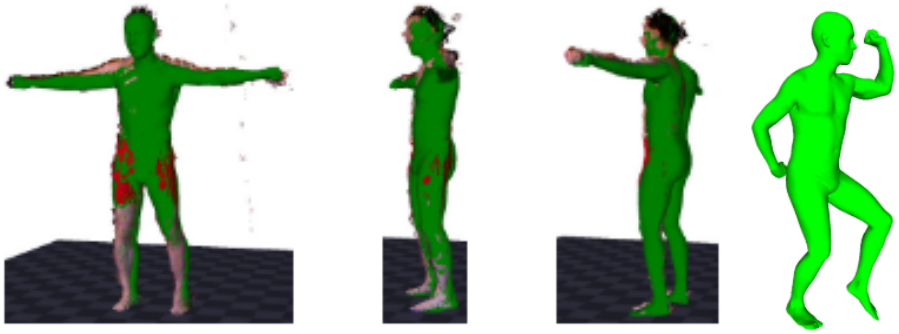
Noisy monocular image and range data was used to reconstruct a parametric invariant to pose model of the human body in [101]. The information was acquired with an RGBD sensor Kinect Xbox 360. The approach couples low-resolution image silhouettes with coarse range data. This method requires the user to rotate in front of the sensor in order to acquire images and depth maps from different views. Since the user is in motion, the body shape changes; thus, the system cannot be treated as a rigid scene. With the aim of tackling this issue, the authors use an a-priori model. Figure 3.6 shows some results obtained with this method.

3.3.2 Reconstruction by fusion: KinectFusion

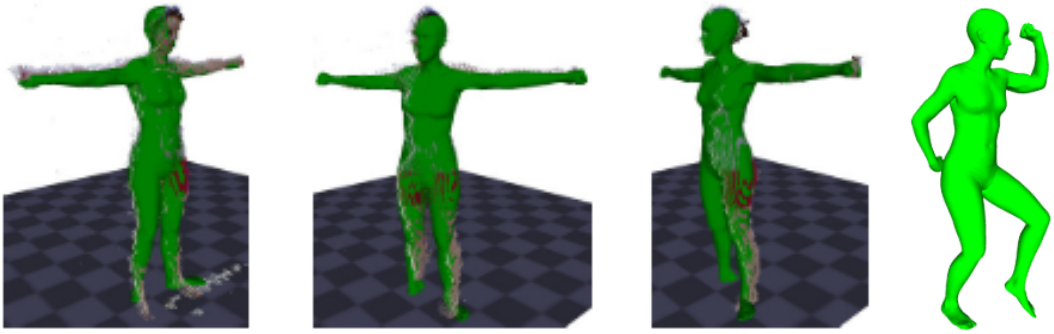
The methods described in this section are 3D reconstruction algorithms of rigid and non-rigid objects. However, they are not based on feature extraction and matching as the approaches described above, they are based on the direct fusion of RGBD scans.

KinectFusion was developed by Newcombe et al. [70, 48] around 2011. This method was developed to work with the RGBD sensor Microsoft Kinect. The algorithm does not use the color information given by the sensor. Thus, it can work on different lighting conditions. The entire reconstruction process runs parallel on a GPU in real time and can be divided into four parts:

1. Depth map conversion: each CUDA thread works in parallel on each pixel, converting the measured depth from the depth map to 3D vertices in camera coordinates. The set of vertices is called vertex map, which is used to create a corresponding normal map using neighboring reprojected points.



(a) Male sample.



(b) Female sample.

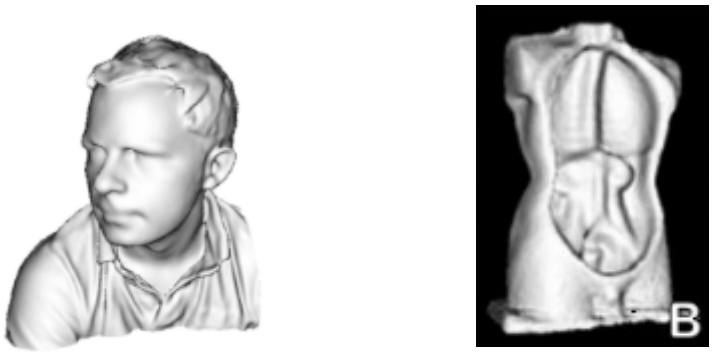
Figure 3.6: The fitted model is represented in green on top of some of the used scans. The model on light green is the fitted model in a different pose. Source: [101].

2. Camera tracking: an iterative closest point (ICP) algorithm is used to track the camera pose for each new depth frame by estimating a single relative six degree-of-freedom (DOF) transformation. Such transformation aligns the current oriented points with those of the previous frame. This process is used incrementally to generate a single global camera pose. Projective data association is used to find correspondences between frames due to the assumption of small camera motion from frame to frame. It is important to recall that during the camera tracking no scene features are detected, the algorithm works on the complete depth maps acquired by the sensor.

3. Volumetric integration: a 3D physical space is represented as a 3D volume with a predefined resolution given by a uniform grid of voxels. The 3D vertices, once mapped to a global coordinate space, are integrated into the voxels considering a Truncated Signed Distance Function (TSDF), which specifies a relative distance to the actual surface. Positive values represent the volume in front of the surface, negative behind the surface, and the surface itself is determined as the zero crossing.

4. Raycasting: a GPU thread works with a single ray. Therefore, taking the starting point and direction of each ray, each GPU thread computes the position of the surface taking into account the change of the sign of the TSDF along the ray. The final position is obtained using linear interpolation. The surface normal is calculated as the derivative of the TSDF at the zero-crossing. Additionally, the calculated vertices and normals are used to compute lighting conditions to render the surface. Furthermore, the raycasted surface can also be used as a less noisy synthetic reference depth map to improve the camera tracking.

Although the approach was developed to reconstruct static scenes, besides tracking the camera motion, the algorithm can also work tracking the six DOF pose of a rigid object rotating in front of a still sensor; thus, creating a second reconstructed volume. Figure 3.7 shows an example of a static scene and a rotating object reconstruction. Additionally, this algorithm can handle small and short motions when reconstructing; due to the dense ICP algorithm used during camera tracking such motions are ignored. However, larger and longer motions as well as fast camera movements might produce a failure in the tracking process.



(a) Static scene.

(b) Reconstruction of an object that was rotated in front of a still sensor.

Figure 3.7: Examples of 3D reconstructions generated in real-time by KinectFusion. Sources: [70, 48].

Several other approaches have been developed either similar to or based on KinectFusion, here some of them are mentioned:

- KinFu is an open-source implementation available in the point cloud library PCL [8]. This implementation can be used online with different RGBD sensors besides the Microsoft Kinect. Furthermore, offline reconstructions are also allowed, such reconstructions only require a set of point clouds and the intrinsic parameters of the sensor used to record them. An implementation for large environments (room size) is also available. The outcome of the algorithm can be stored as point cloud

and mesh in different formats like .ply and .vtk. The resulting reconstructions can be texturized if the user requires it.

- ReconstructMe [10] is a software, very similar to KinectFusion, developed by PROFACTOR GmbH. The software offers different types of licenses, such as non-commercial user, commercial user, and commercial SDK developer. The free developer license comes with some usage limitations: low mesh generation, mesh and image watermarks. ReconstructMe offers scanning of different volumes and has a selfie mode, from which a printable-color 3D bust (watertight, scaled, able to stand) is the outcome.

- CopyMe3D is an approach developed by Sturm et al. [94] to generate 3D people color reconstructions, these models are very similar to the ones given by ReconstructMe in selfie mode. Furthermore, this method presented more effective tracking and weighting approaches in comparison to KinectFusion's.

- DynamicFusion [69] can be seen as the evolution of KinectFusion. This algorithm reconstructs non-rigid deforming scenes in real time and allows camera and scene motion at the same time. The algorithm transforms the state of a scene at each time into a fixed, canonical frame. For example, if the scene is a person in motion, the person's motion is removed warping each body configuration into the pose of the first frame; then, the scene is rigid and regular KinectFusion updates can be used to generate the 3D reconstruction. Such reconstruction is transformed back to the live frame in order to show the 3D reconstruction of the moving scene, see Fig. 3.8.

Figure 3.8: Live reconstruction of "drinking from a cup" scene generated by DynamicFusion. Source: [69].

- Other KinectFusion related works are found in [86, 74, 61, 53, 17].

Chapter 4

Head and Ear Measurements

Anthropometric dimensions are used in different fields, such as medicine, acoustics, and ergonomics. This chapter describes the automatic estimation of head and ear measurements from 3D reconstructions (uncolored point clouds) of the upper part of the body. Such measurements are

- head width,
- head height,
- head depth,
- the closest distance from the front of the head to the ear pits (distance between the nose bridge and the ear pits),
- the closest distance from the back of the head to the ear pits (distance between the nape and the ear pits), and
- helix height.

These measurements were estimated due to the requirements to personalize head related transfer functions (a person-dependent function to locate sound sources). However, the measurements can be used in other applications such as design of helmets, masks, and headphones, among others. In order to estimate the aforementioned dimensions, it is necessary to locate feature points that act like reference points. These points are the nose tip, the ear pits, the top of the head, the nose bridge, the nape, the chin, the forehead, and the most prominent part at the back of the head.

Previous work related to head and ear measurements and to pure feature points detection is presented in Section 4.1. The proposed automatic approach to locate the abovementioned features points and to estimate the measurements is described in Section 4.2. In order to evaluate the described approach, experiments and results are presented in Section 4.3. Finally, a summary and conclusions are given in Section 4.4.

This chapter is based on the paper: *Head measurements from 3D point clouds* authored by Patiño and Zell [76]. Some fragments presented here are replicated from the original paper.

4.1 Related Work

4.1.1 Head and ear measurements

Anthropometric dimensions, including head measurements have been used to calculate the anthropometric variations in Europe and Mediterranean area in [33]. The authors collected measurements as head maximum length and breadth of 85 men. Farkas et al. studied the growth of the head in 1537 North American Caucasians (see [38]). The head width (as the distance between the eurions), height (measured between the nose bridge and the top of the head), depth (calculated between the glabella and the opisthocranion), and circumference, as well as the forehead width were directly measured. In order to develop a sizing system for protective clothing of the New Zealand fire services, measurements of 691 men were directly obtained. The acquired anthropometric data included head depth and width (see [56]). Additionally, the head depth and width (between the eurions) were measured with a spreading caliper in a study of mixed-race USA Army male soldiers for military design sizing in [110] (other measurements were obtained using an automated headboard device). Another study manually measured the head width (measured between the gonions), among others, to test and design respirators (see [111]).

The Caesar project (see [85, 84]) collected anthropometry information of people in North America (USA and Canada), the Netherlands, and Italy. The measurements were obtained using traditional tools and 3D models acquired with a Cyberware WB4 and a Vitronic scanner. Before capturing the 3D models, 72 landmarks were placed on the subjects. The head width and depth were measured with traditional tools. However, other measurements as the distance between the right and left tragus were measured on the 3D models. The extraction of the 3D location of the landmarks was done semi-automatically. The CIPIC database contains anthropometric information, including head width and height, as well as ear measurements (pinna height and width) of 43 subjects obtained with high resolution photographs, measurement tape, and a Polhemus 3D stylus digitizer (see [19], [18]). Additionally, body and ear dimensions, as torso radius and pinna height were directly acquired using photographs of nine subjects in [112]. Rothbucher et al. obtained anthropometric dimensions from 3D models acquired with a system based on triangulation. Such dimensions were manually measured on the 3D models using MeshLab (see [87]). Head and ear measurements (the same ones computed in the CIPIC database) were acquired in [97] using an automatic process based on active shape models on 2D images. The authors worked with pictures of 15 subjects obtained under a controlled setup. Dinakaran et al. [35] automatically estimated head measurements from 3D head-shoulder meshes of 40 subjects obtained with a Kinect sensor. KinectFusion was used to generate the models. However, manual post-processing was performed. Measurements as the head width, depth and height, as well as pinna height and depth were obtained evaluating the outlines of the model. Furthermore, during the generation of the ITA database [25], head and ear measurements of 48 subjects were collected. Measurements such as the head height (measured from the center of the

head to the top) and the pinna height were obtained from magnetic resonance images and 3D ear models.

4.1.2 Face features and ear detection in 3D

Segundo et al. proposed an algorithm to detect the nose tip and base, as well as nose and inner eye corners in range images (see [90]). The authors combined surface curvature analysis with 2D facial feature techniques applied to 3D, such as projection of the topographic depth information relief. Perakis et al. [79] presented a method to automatically detect eye and mouth corners, nose and chin tips from 3D facial scans. The authors used local shape descriptors: shape index and spin images to extract candidate landmark points. The landmarks were identified and labeled by matching them with a facial landmark model, which was created during the training phase. A voting framework with patches extracted from the whole depth image was proposed by Fanelli et al. to detect facial features and estimate the head pose in [37]. Such features as eye, nose and mouth corners were localized using Random Forests [28]. The Discriminative Generalized Hough Transform was used in [24] to detect facial landmarks in 3D meshes. The localization of the features was performed in consecutive steps increasing the grade of detail and narrowing down the region of interest.

The ear pit was found automatically through skin detection, curvature estimation, and surface segmentation and classification in a 3D + 2D profile image combination in [107]. Additionally, an active contour algorithm was used to localize the ear outline. Moreover, an automatic scale and rotation invariant ear detection technique was proposed by Prakash and Gupta in [80]. The ear was detected on profile face range images using an edge connectivity graph and a template. Ding et al. [36] used side range images to detect the ear. The ear region was isolated based on the location of the nose tip. The ear pit was located under the assumption that the pit was the point with the lowest z value in a predefined circle window. If several candidates arose, the final pit location was defined as the place which contained more candidates. The ear contour was also found using a manually obtained template. Lei et al. proposed an ear tree-structured graph to represent a 3D ear on range side facial images in [58]. The authors defined the tree with 18 landmarks on the helix and anti-helix regions, choosing the beginning of the helix as the root of the tree. A score for each root candidate was calculated taking into account the best possible positions of all the landmarks. The root with the highest score determined the location of the ear.

4.2 Feature Points Estimation and Measurements

In order to obtain measurements, it is necessary to estimate feature points. Therefore, with the aim of measuring the head width, the ear pits are required. The head height is measured from the top of the head to the chin. The depth of the head is the distance

between the forehead and the back of the head. Other feature points as the nose bridge and the nape are also required with the purpose of calculating the distance from these points to the ear pits. To estimate the pinna height, only the ear pit is needed. Such feature points and measurements are depicted in Fig. 4.1.

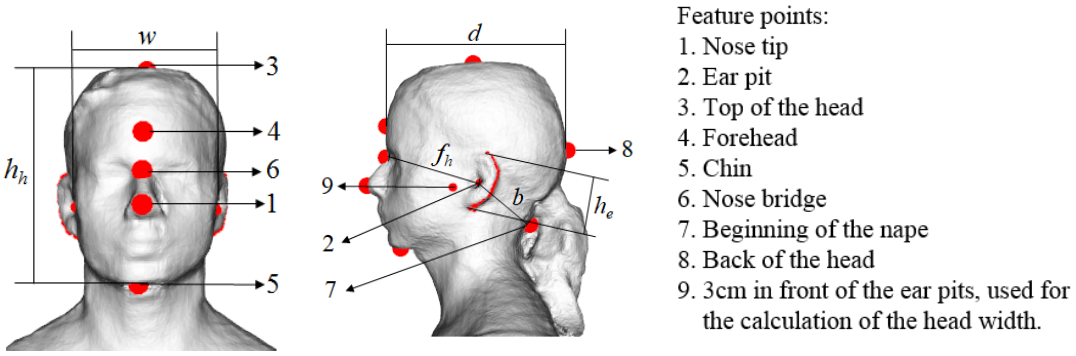
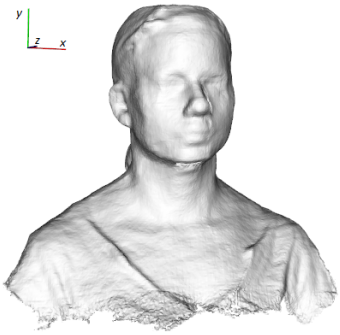


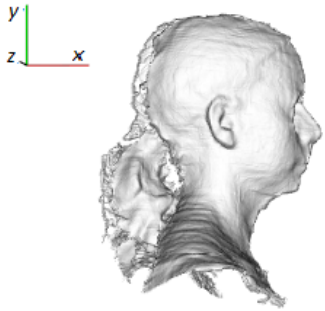
Figure 4.1: The features required to obtain head and ear measurements as well as the measurements are depicted. The measurements are labeled as follows: w stands for the head width, h_h is the head height, h_e describes the helix height, d represents the head depth, f_h is the shortest distance from the ears to the front, and b symbolizes the shortest distance from ears to the back.

In this section, the approaches to automatically locate the aforementioned feature points and to estimate the measurements are described. Three 3D point clouds, shown in Fig. 4.2, of the upper part of the body per subject were used: two profile reconstructions (right and left), and one 360° (complete) reconstruction. For the acquisition of the two profile point clouds, both the sensor and the subject remained still. On the other hand, for the obtainment of the complete point cloud, the sensor remained still while the subject rotated in front of it. Furthermore, since the measurements are estimated on the complete point cloud, the profile reconstructions are used as auxiliary data.

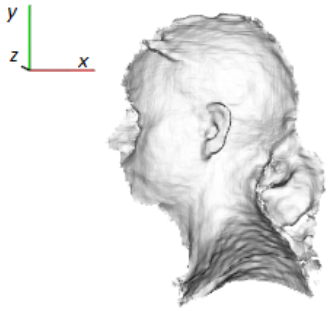
The sequence of the feature points estimation is depicted in Fig. 4.3, this is the order in which the algorithm was designed but it does not always represent the requirements to locate each point. However, since the ear pits are used to align the head to the world coordinate system, they represent the base of the entire algorithm and their positions should be estimated before any other feature point, except the nose tip. Furthermore, although the helix is not a point, it is included in the diagram as the final step of the process since it is a feature needed to estimate the pinna height. Table 4.1 shows the direct dependency of each feature point and the helix, i.e. the points directly required to estimate each specific feature point.



(a) Complete (360°) reconstruction.



(b) Right profile reconstruction



(c) Left profile reconstruction.

Figure 4.2: 3D reconstructions used to automatically estimate head dimensions. The reconstructions were acquired with a sensor Asus Xtion Pro Live and the reconstruction algorithm KinFu. The axes representation is x, y, z for red, green, and blue lines, respectively. The origin of the axes does not correspond to the actual origin of the coordinate systems.

4.2.1 Ear pits estimation from the profile point clouds

The complete reconstruction tends to have imbalanced quality since the motion of the subject affects the reconstruction. Such unevenness might be exhibited as differences in the ears. Therefore, the parameters used to find the ear pit on the right ear may not work for the left ear. Hence, a dynamic approach is needed. In order to be able to use a dynamic approach to detect the ear pits on the complete point cloud, a small region of interest (ROI) around each ear has to be established. This small ROI is obtained based on the location of the ear pits on the profile reconstructions, which present the same quality since the subjects do not move during their acquisition.

In order to estimate the location of the ear pit on a profile reconstruction, it is necessary to determine a region of interest around the ear. According to the literature [106], the ear region can be determined based on the nose tip position. Therefore, after removing the noise and outliers from the profile point clouds, the nose tip is found as the rightmost

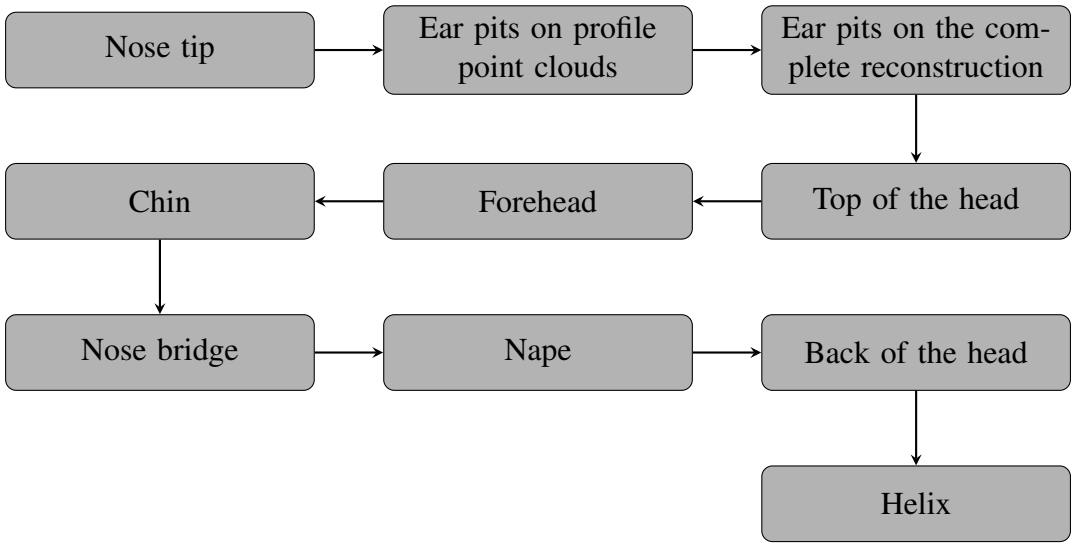


Figure 4.3: Process diagram of the estimation of feature points and the helix. The sequence indicates the order in which the algorithm was designed, it does not represent the requirements to estimate each point.

		Requirement									
		Nose tip	Ear pits (profile)	Ear pits (complete)	Top of the head	Forehead	Chin	Nose bridge	Nape	Back of the head	Centroid
Feature point	Nose tip	-	-	-	-	-	-	-	-	-	-
	Ear pits (profile)	✓	-	-	-	-	-	-	-	-	-
	Ear pits (complete)	✓	✓	-	-	-	-	-	-	-	-
	Top of the head	-	-	-	-	-	-	-	-	-	-
	Forehead	✓	-	-	✓	-	-	-	-	-	-
	Chin	✓	-	-	-	-	-	-	-	-	-
	Nose bridge	✓	-	-	-	✓	-	-	-	-	-
	Nape	✓	-	-	-	-	-	-	-	-	✓
	Back of the head	-	-	-	-	-	-	-	✓	-	-
	Helix	-	-	✓	-	-	-	-	-	-	-

Table 4.1: Direct requirements to estimate the location of each feature point.

or leftmost point above the centroid of the reconstruction. There are cases in which the hair style affects the nose tip detection in one of the profiles. Hence, a comparison of the y coordinate of the tips on both the right and left profile is performed. In case the y value of both nose tips varies more than a determined threshold, the lowest y value is selected to create a ROI on the point cloud with the highest nose tip. In such ROI either the rightmost or leftmost point is detected and categorized as the nose tip of that specific profile point cloud. The nose tip with the lowest y value is not modified.

The ear is likely to be inside a rectangular area (9×11 cm) that starts at 7 cm from the nose tip along the horizontal axis, as shown in Fig. 4.4. Inside the determined ROI, the ear pit lies on a concave area. Therefore, a concavity study is performed (it is inspired by [105]). Considering that the cosine of the angle between two vectors tends to be positive if the intersection of the vectors lies on a concave area, negative if the vertex lies on a convex area, and zero if the junction lies on a plane, each point in the ROI is analyzed with regard to its neighborhood K using the equation

$$cos\,\theta_j = \frac{\vec{n}_{p_i} \cdot (\vec{p}_j - \vec{p}_i)}{\|\vec{n}_{p_i}\| \, \|\vec{p}_j - \vec{p}_i\|} \,, \tag{4.1}$$

where $\vec{p}_i$ represents the interest point, $\vec{n}_{p_i}$ is its normal, $\vec{p}_j$ describes the $j-th$ neighbor, and $\cdot$ stands for dot product. A graphical illustration is shown in Fig. 4.5. The neighborhood K_n consists of all the points inside a sphere of radius r centered at $\vec{p}_i$. In order to guarantee that $\vec{p}_i$ lies in fact on, for example, a concave area, it is necessary that the cosine of the angle between its normal and the link to most of its neighbors is positive. However, small positive and negative values represent almost-planar sections; thus, thresholds are recommended and used.

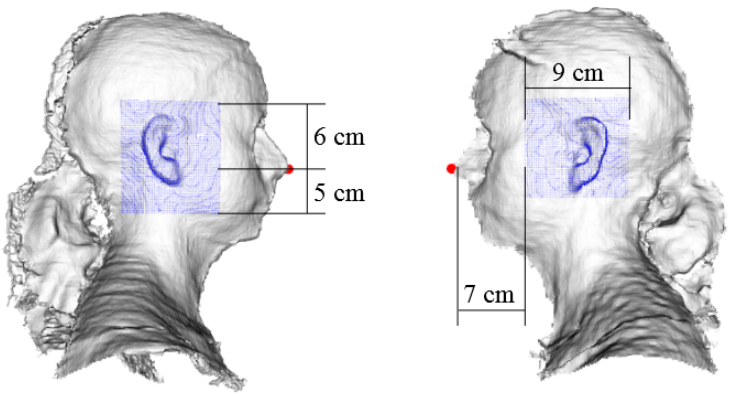


Figure 4.4: Region of interest around the ear on the profile reconstructions. The red point identifies the detected nose tip. The blue area is the ROI obtained based on the nose tip position.

To estimate ear pit candidates, two point clouds are generated in the ROI:

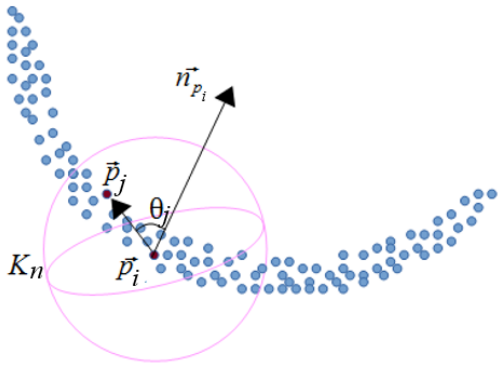


Figure 4.5: Concavity study. The cosine of the angle between two vectors is positive if their intersection lies on a concave area. Therefore, the point $\vec{p}_i$ lies on a concave area if the cosine of the angle between its normal $\vec{n}_{p_i}$ and the link to its neighbor $\vec{p}_j$ is positive. All the points inside a sphere centered at $\vec{p}_i$ belong to the neighborhood K_n. (Image inspired by Fig. 1 in [105])

1. *concaveCloud* that contains points lying on concave areas and,
2. *noAreaCloud* that consists of points that do not lie on convex, planar, or concave areas.

The candidates are points that belong to the point cloud *concaveCloud* that are surrounded by a determined number of points of *noAreaCloud*. The exact ear pit is estimated as the point surrounded by more pit candidates. In Fig. 4.6 the point clouds *concaveCloud* and *noAreaCloud* are represented by the green and red points respectively, the purple points are the ear pit candidates, and the ear pit is demarcated by the cyan point (the aforementioned points are enlarged for rendering purposes).

4.2.2 Ear pits estimation from the complete (360°) reconstruction

Since the position of the ear pits found on the profile point clouds do not always correspond to the ear pits on the complete reconstruction (due to e.g. the user's alignment and displacements), regions of interest have to be established with the aim of estimating the ear pits on the complete point cloud. These ROIs are determined based on the coordinates of the ear pits found on the profile reconstructions. Therefore, it is necessary to set a common feature on the profile and complete point clouds to work as a reference point. Since the nose tip is the most salient feature on the head, it is chosen as the reference point. Considering that the initial position of the user, when recording the complete cloud, is looking at the sensor, the nose tip is defined as the closest point to the camera.

Once the coordinates of the ear pits estimated on the profile point clouds are calculated with respect to the nose tip of the same point clouds, they are used as center points of

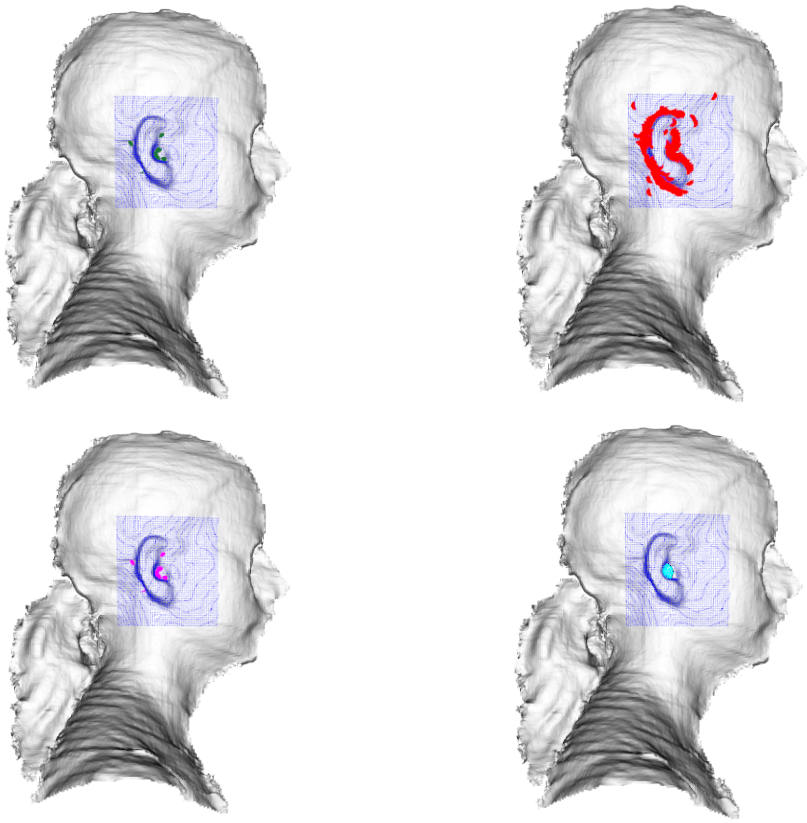


Figure 4.6: In order to estimate the ear pit from profile point clouds, a concavity analysis is performed. The blue area represents the ROI, the green areas describe the location of points lying on concave areas (*concaveCloud*), points that do not belong to concave, convex, or planar areas are shown in red (*noAreaCloud*), the purple points are the ear pit candidates, and the cyan point represents the estimated ear pit.

the ROIs at each side of the complete reconstruction. These center points are measured with respect to the nose tip estimated on the complete cloud. Each ROI is a rectangle of 4×10 cm. However, in order to avoid hair and neck areas, the regions of interest are vertically limited to not exceed the nose tip position by 3 cm and -4 cm. The process to calculate the ear pits position on the head varies in comparison to the one explained in Section 4.2.1 in five aspects:

1. more than two ear pit candidate neighbors are required to select a point as the ear pit,

2. if the ear pit is not found, the threshold used to classify a point as lying on a concave area iteratively decreases. The decrement of the threshold means an increment of the angle of aperture of the possible concave area,

3. in case that the approach described in 2. does not find the ear pit, the threshold used to classify a point as on a concave area is reset, and the radius of the neighborhood used to look for ear pit candidate neighbors increases,

4. supposing that two points are selected as the ear pit because they have the same amount of candidates neighbors, the distance of the possible ear pit to these neighbors is taken into consideration. The candidate with the shortest distance to its neighbors is recognized as the ear pit,

5. with the purpose of reducing false positives on the helix and earlobe, these regions are filtered out on the point cloud *noAreaCloud*.

4.2.3 Data preparation for further use

The orientation of the complete point cloud is affected by the inclination of the sensor and the posture of the user. Therefore, the reconstruction is not always aligned to the vertical axis of the real-world coordinate system, on occasions even the head and the torso have different orientations. Since the position of the feature points depends on the alignment of the head, Principal Component Analysis (PCA) is applied to the upper part of it (from the nose tip to the top) to compute its real orientation. Additionally, after PCA, a refined alignment is performed based on the location of the ear pits. Such a refined alignment consists in the placement of the ear pits at a similar height and parallel to the x axis. Figure 4.7 presents the comparison of a head reconstruction before and after the alignments.

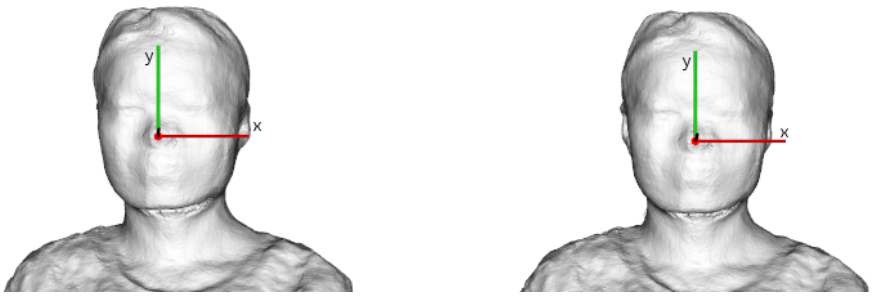


Figure 4.7: Principal component analysis (PCA) and refined alignment. Right: the original reconstruction. Left: the reconstruction after PCA and refined alignment. The axes representation is x, y, z for red, green, and blue lines, respectively.

4.2.4 Top of the head, chin, nape, and back of the head feature points estimation

The estimation of some head feature points as the top and back of the head, the chin, and the nape are affected by hair and facial hair styles. However, the hair and facial hair (from now on both called hair for simplicity) are part of the user's physical appearance reconstructed by KinFu. Therefore, the measurements obtained with the approaches explained in this chapter include the hair.

The top of the head is estimated as the highest point of the complete point cloud. For rendering purposes, the point is depicted parallel to the ear pits along z and at the origin of the x axis.

In order to estimate the chin feature point, which is located at the end of the face (see Fig. 4.1), the chin region is isolated through a rectangular area (2×3 cm) located 7 cm below the nose tip. On this region, the closest point to the camera, called *salientChinPoint*, is determined with the aim of reducing the search area. Such reduced ROI contains the points located below the *salientChinPoint* and at a distance not larger than 1.5 cm along the z axis from the mentioned point (see Fig. 4.2 to recall the coordinate system). A relaxed concavity analysis is performed on the new determined region of interest. It is called relaxed because the condition to determine a point as lying on a concave area (positive result of eq. 4.1 for all the neighbors) is relaxed since the chin is not as concave as the ear pit. Now, the interest point only needs a positive angle with more than ten neighbors to be considered as lying on a concave area. Finally, the chin feature point is determined as the candidate lying on the most concave area. The chin estimation process is depicted in Fig. 4.8.

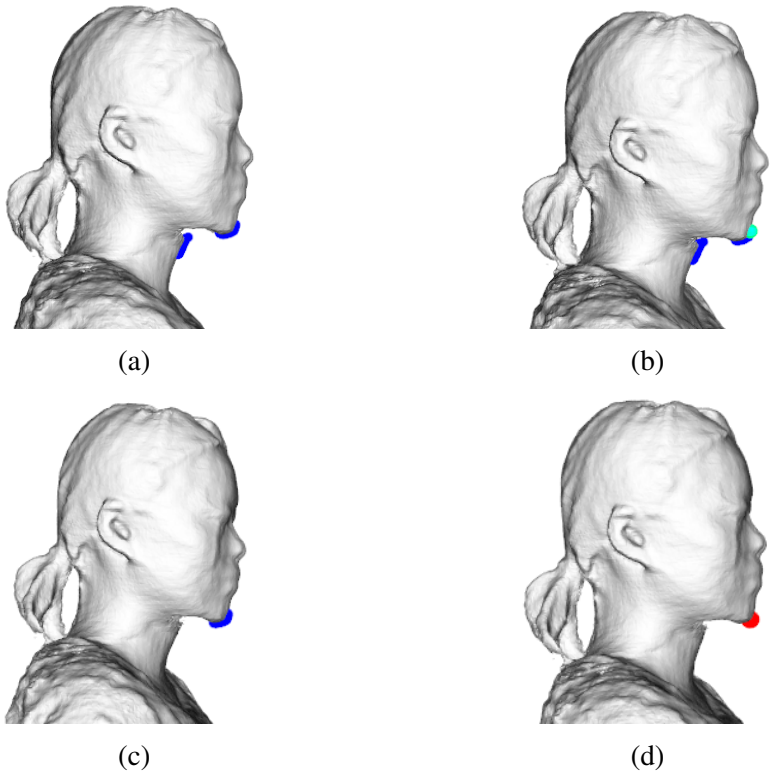
The nape is the closest point on the back of the head to the ear pits. Therefore, a region of interest is set, after filtering out outliers, with points of the back of the head, located 6 cm below the top of the head, but above the centroid of the complete reconstruction, and at the middle of the head (half way the horizontal path - along the x axis - between the ear pits). The location of the nape is estimated through the minimization of the distance between the ear pits and each point of the ROI. Since there are two ear pits, the average of the distance of each ear pit to the point of the ROI is the one minimized. Figure 4.9 shows a graphical illustration of the aforementioned process.

The back point of the head is the furthest point to the camera 5 cm or more above the nape located at the middle of the head (half way the horizontal (along the x axis) path between the ear pits).

4.2.5 Forehead and nose bridge feature point estimation

The forehead is estimated as a point half way to the top of the head from the nose tip, with the same horizontal coordinate than the nose tip.

The nose bridge is the closest point on the front of the head to the ear pits. Therefore, a region of interest is established, after filtering out the outliers, with points of the front of



(a) (b)

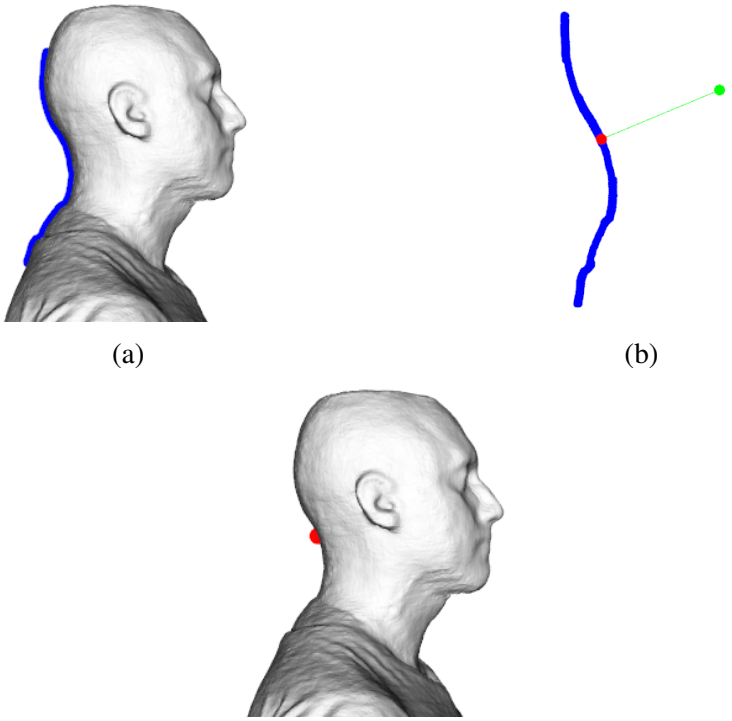
(c) (d)

Figure 4.8: Chin estimation process. (a) The first region of interest is colored in blue. (b) The cyan sphere represents the *salientChinPoint*. (c) The second region of interest is blue. (d) The chin feature point is shown in red.

the head located 1 cm above the nose tip, but below the forehead, with the same horizontal coordinate (x) than the nose tip. As done for the estimation of the nape location, the nose bridge position is calculated through the minimization of the distance between the ear pits and each point of the ROI. Since there are two ear pits, the average of the distance of each ear pit to the point of the ROI is the one minimized. A graphical illustration of the explained process is shown in Fig. 4.10.

4.2.6 Helix detection

A ROI of 7×10 cm is created on the complete model around the ear pits in order to detect the helix. Since the 3D models do not present any space between the pinna (external part of the ear) and the head, a 3D edge detector is used to estimate the helix. In such an edge detector, the region of interest is divided into horizontal and vertical lines. This is done with the purpose of analyzing the neighborhood of each point on each line looking for depth differences (the ROI is studied horizontally and vertically independently). The depth differences determine the points that belong to 3D edges. Such points are collected



(a) (b)

(c)

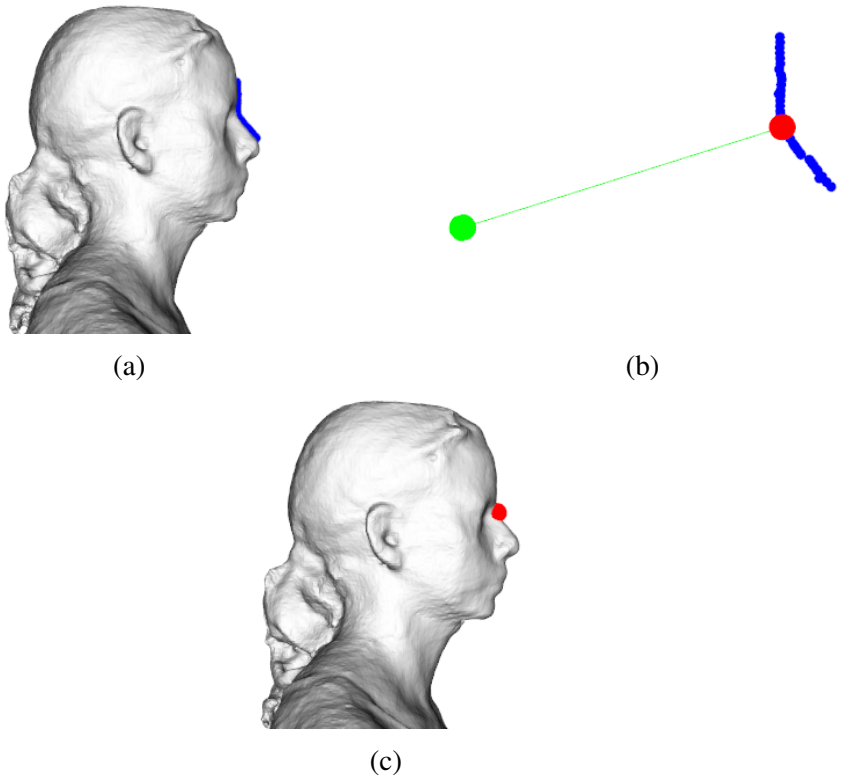
Figure 4.9: Nape estimation. (a) The region of interest is colored in blue. (b) The ear pit is represented as a green sphere, the line connects the ear pit to the closest point (red dot) in the ROI. (c) The nape is indicated by the red sphere.

on a 3D edges point cloud. Afterwards, points with small curvature as well as points belonging to small clusters are filtered out from the 3D edges point cloud. From the general geometry of the ear, it is possible to deduce that the helix is its outermost part. Therefore, the point cloud with the 3D edges is divided into horizontal and vertical lines with the aim of finding the outermost point of each line (the horizontal and vertical studies were performed independently). The set of outermost points represents the helix candidates.

An analysis of the helix candidates' position is conducted with the aim of selecting the points that belong to the helix:

1. the first point on the helix is selected from the outermost points found on the horizontal lines. It is the first point that allows a smooth transition in ear-shape to the next outermost horizontal point,

2. in order to determine the subsequent points that belong to the helix, the neighborhood of the last point on the helix is analyzed. Points that are close to the last helix point are considered as part of the helix as well,



(a) (b)

(c)

Figure 4.10: Nose bridge estimation. (a) The region of interest is colored in blue. (b) The ear pit is represented as a green sphere, the line connects the ear pit to the closest point (red dot) in the ROI. (c) The nose bridge is indicated by the red sphere.

3. in case the last helix point does not have any neighbors, the ear is divided in three equal horizontal sections called *upper-*, *middle-*, and *lower-ear*. The helix follows a specific direction in each of these sections; therefore, the volume used to search for points member of the helix depends on the section at which the last helix point belongs to. Such search volume is larger than the one used in 2. The closest point to the last helix point found in the new search area is taken as part of the helix.

A graphical example of the 3D edge detector, the helix candidates, and the estimated helix is depicted in Fig. 4.11.

4.2.7 Measurements

The head width is estimated as the horizontal distance (along the x axis) between two points located 3 cm in front of the ear pits. These points are close to the cheek bones. The head height is the vertical distance between the top of the head and the chin. The horizontal distance (along the z axis) between the forehead feature point and the back

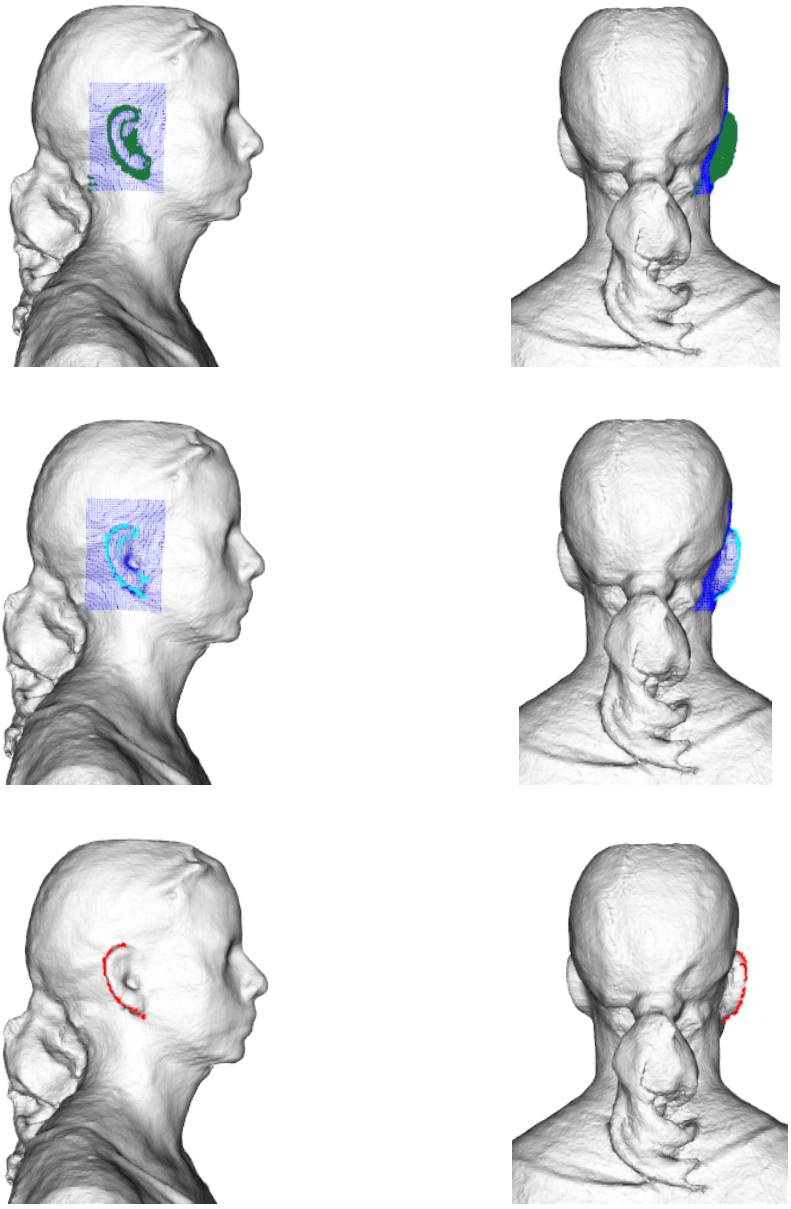


Figure 4.11: Helix detection. Side and back views of a subject are shown. The blue area represents the ROI, the green areas describe the 3D edges, the cyan points symbolize the helix candidates, and the helix is depicted in red.

of the head are used to compute the head depth. The distances from the ear pits to the nape and the nose bridge are already established in Section 4.2.4 and Section 4.2.5, respectively. The height of the helix is measured as the difference between the uppermost and the lowest point on the helix on the plane *yz*, see Fig. 4.1 and Fig. 4.2 for a graphical

description of the helix height and the coordinate system, respectively.

4.3 Experiments and Results

In order to test the measurements estimation method explained in this chapter, the 3D reconstructions of 20 subjects were captured. Considering the great performance of the RGBD sensors and the promising results given by fusion algorithms such as KinectFusion, the data were recorded using an RGBD sensor Asus Xtion Pro Live (state of the art at the time) and the software KinFu (since it allows to use other sensors than the Microsoft Kinect). For each subject, four sets of 3D reconstructions were acquired: 1. the subject rotated in clockwise direction, and the reconstruction was performed at $\approx$ 5 fps, 2. the subject rotated in counter-clockwise direction, and the reconstruction was performed at $\approx$ 5 fps, 3. the subject rotated in clockwise direction, and the reconstruction was performed at $\approx$ 10 fps, and 4. the subject rotated in counter-clockwise direction, and the reconstruction was performed at $\approx$ 10 fps. Each set contained two profile reconstructions (right and left side) and a complete (360°) reconstruction. The data were used to evaluate:

1. the ear pit detection on the complete reconstruction by means of a visual analysis,
2. the accuracy of the estimated measurements through the comparison of the estimated values with ground-truth data,
3. the reproducibility of the designed method using the four sets of reconstructions per subject to compare the estimated values under different data acquisition conditions.

The evaluated measurements were the head width, height, and depth. The estimation of the position of the nose bridge and the nape was evaluated as the distance from these points to the top of the head. The top of the head was used instead of the ear pits because the ear pits ground truth location was not feasible to acquire.

Large figures of this section are shown at the end of the chapter to avoid extensive interruptions in the text.

4.3.1 Data acquisition

The 3D data were acquired online with an RGBD sensor Asus Xtion Pro Live and the KinFu application available in PCL. Two graphic cards were used: a Nvidia GeForce GT 630 and a Nvidia GeForce GTX 560 Ti. Both graphic cards have a memory of $\approx$ 1 GB. However, the GT 630 builds online 3D reconstructions at $\approx$ 5 fps with resolution of 512 voxels per axis (vpa), whereas the GT 560 Ti produces online 3D reconstructions at $\approx$ 10 fps with a resolution of 544 voxels per axis.

Since the 3D reconstructions obtained with KinFu are the representation of the current physical appearance of the subjects, people to be scanned should fulfill some requirements:

- hats, large earrings and piercings are forbidden,
- the hair should be away from the ears, particularly from the ear pits,
- at the front of the face, the hair should not go beyond the nose tip,
- if the user has long hair, a low ponytail is required. However, if the ponytail obstructs the nape, a hair bun is needed. It may be the case, that two complete reconstructions are necessary (one with a ponytail and other with a bun) in order to obtain the nape and back of the head feature points.

With the aim of obtaining the reconstructions, the sensor was intended to be parallel to and at the eyes' height of the subjects. The person was located at ≈ 0.58 m (mean distance among the 20 subjects) from the camera. The working (reconstructed) volume was limited to $1\,\mathrm{m}^3$. In order to record the complete reconstruction, the subject was placed looking at the sensor, then he/she slowly rotated until a revolution was fulfilled. During the rotation, pauses of ≈ 10 seconds were done at $\approx 90°$ and $270°$ to improve the ear data acquisition. To obtain the profile reconstructions, the subject stood still showing the right or left side of the head and body to the sensor. Four reconstruction sets per each of the 20 subjects were acquired at different frame rates when each subject rotated in different directions (as explained before). The age of the subjects was between 20 and 35 years. Three out of the 20 persons were female. Males with different amount of face hair were included. Furthermore, small earrings and piercings were not removed.

4.3.2 Ground-truth values acquisition

In order to evaluate the accuracy of the proposed method, the ground-truth values were obtained by means of:

1. a caliper and a measurement tape, the caliper was specially designed by the RWTH Aachen, it is depicted in Fig. 4.12. The caliper was used to obtain the head width, whereas the measurement tape was employed to measure the head height and depth, and the distance from the nose bridge and the nape to the top of the head. However, the geometry of the human head impeded the accurate acquisition of measurements with the measurement tape. Therefore, only the width captured with the caliper was used for evaluation,

2. a tracking system, which consisted of six OptiTrack Flex V100 cameras. Eight spherical markers (11.3 mm of diameter) were placed on the head of each subject: two at each side close to the ears to determine the head width, one on top of the head and another on the chin to calculate the height, one on the forehead and one on the back

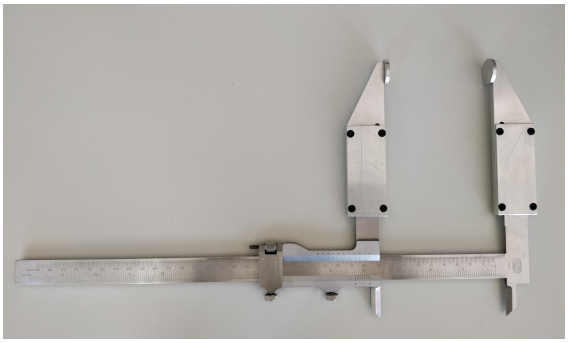
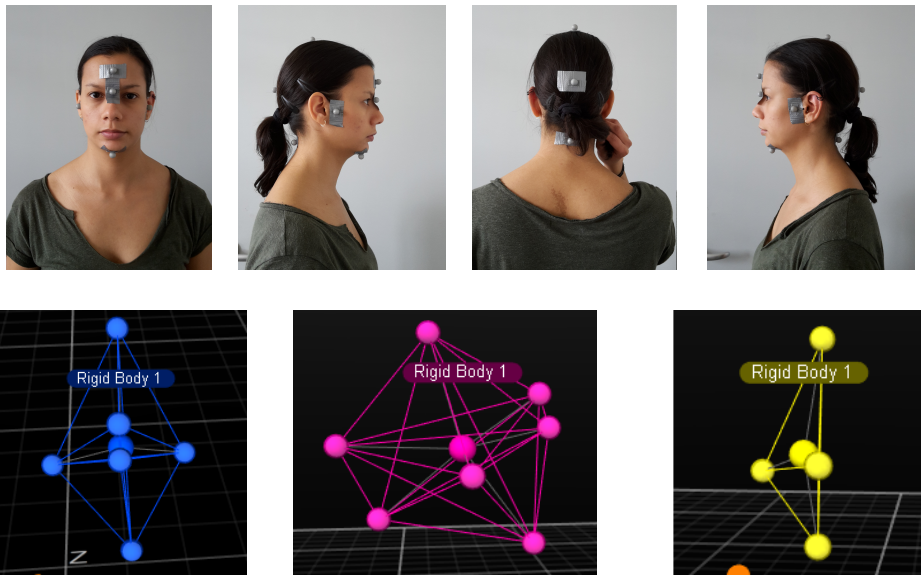


Figure 4.12: Caliper used to manually measure the head width.

part of the head to estimate the depth, one on the nose bridge, and one on the nape to calculate the distance from these points to the top of the head, see the display of the markers in Fig. 4.13. The measurements were obtained based on the position of each marker. For the head width and depth only the horizontal coordinate was considered, whereas for the head height, and the distance from the nose bridge and the nape to the top of the head only the vertical coordinate was utilized. Fig. 4.14 shows the setup of the OptiTrack cameras used to obtain the markers' position on a subject. The cameras are represented as orange pyramids, the markers on top and both sides of the head, on the chin, the forehead and the nose bridge are depicted in blue,



(a) Frontal view. (b) Side view. (c) Back view.

Figure 4.13: Markers used together with the tracking system to obtain ground truth measurements. Top: multiples views of the markers placed on a subject. Bottom: visualization of the markers in the software Motive (Optitrack software).

3. the highest and lowest point of the pinna were manually selected on the mesh to obtain the ground-truth helix height. The software used for this purpose was MeshLab [73], which is a free software to process and edit meshes and point clouds.

4.3.3 Visual evaluation of the feature points location estimation

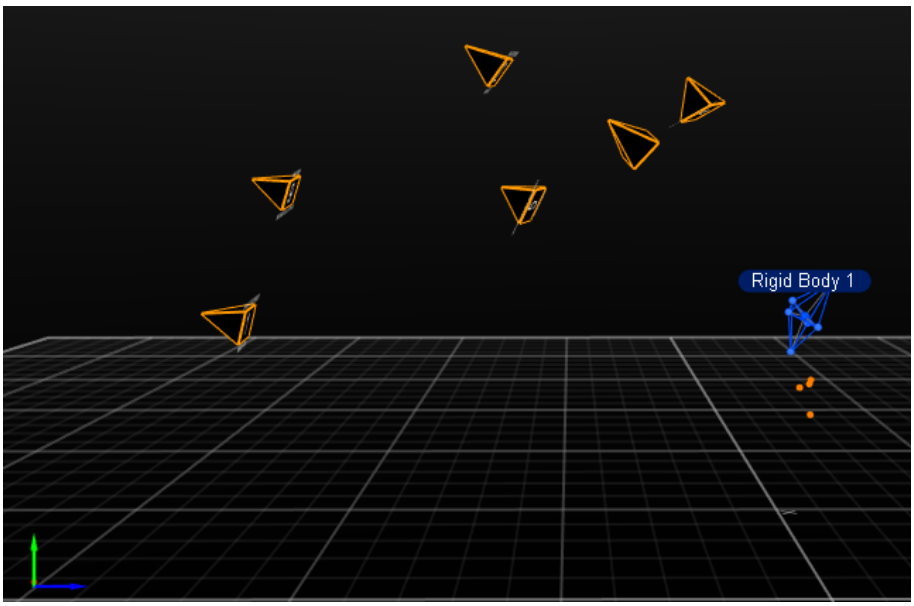
The 3D reconstructions presented holes and artifacts on areas that were not scanned by the sensors, as expected. Such areas are: the top of the head and the space between the neck and the chin. For some subjects, holes and artifacts were also observed behind the ear lobe. Nevertheless, the algorithm presented in this chapter was designed to work with such affairs. Figure 4.15 shows the feature points located on 3D reconstructions of four different subjects, each row represents a different data acquisition condition: first and second rows were reconstructed at $\approx$ 5 fps, with resolution of 512 vpa, while the subject rotated in clockwise and counter-clockwise direction, respectively; third and fourth rows depict reconstructions at $\approx$ 10 fps, with resolution of 544 vpa, during clockwise and counter-clockwise rotations, respectively.

Visually, the 3D models obtained at $\approx$ 10 fps (544 vpa) exhibited a smoother surface than those obtained at $\approx$ 5 fps (512 vpa). However, the detected feature points did not show any divergence between the two type of models.

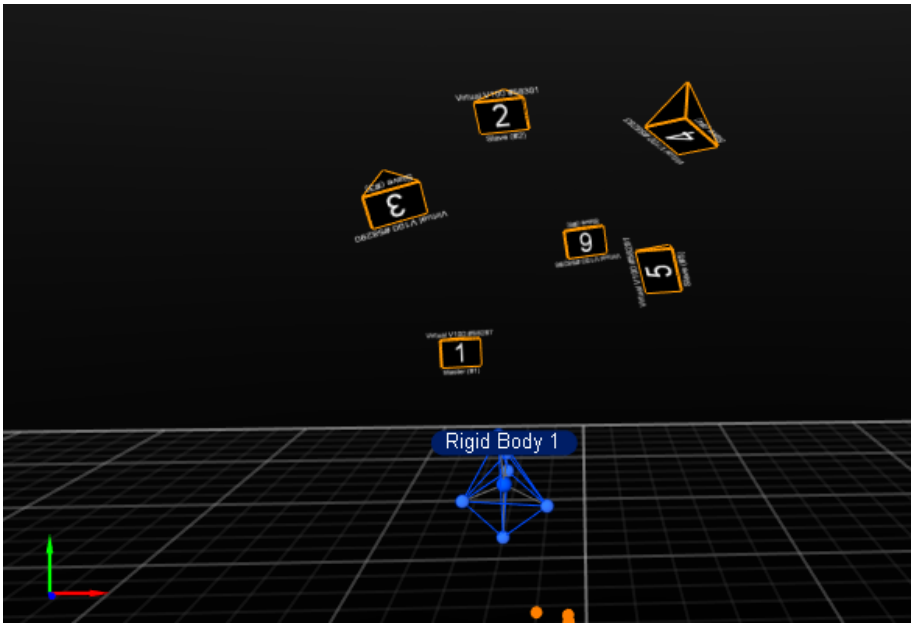
4.3.4 Ear pit detection evaluation

An ear pit was classified as correctly estimated if it was located inside the volume comprehended by the concha, see Fig. 4.16. Such classification was performed visually. The presented evaluation was performed only for the complete reconstructions. The dataset acquired with the process explained in Section 4.3.1 provided 160 ears. However, two samples were excluded from the ear pit detection evaluation because they presented local defects on the point cloud. The results showed a detection rate of 94.3 %; Ding et.al. [36] reported an ear pit detection accuracy of 90.9 %, the approach developed in [106, 107] also detected the ear pit. However, the accuracy rate was not reported. Figure 4.17 exhibit ear pits classified as good and bad estimations, respectively. The bad ear pit localization might be produced by different factors as:

- a bad ear pit localization on the side point clouds. This would generate a bad ROI extraction on the complete reconstruction,
- concave areas produced by the hair,
- artifacts or malformations on the reconstruction,
- a deep triangular fossa.



(a) Side view.



(b) Subject's point of view

Figure 4.14: Setup of the six OptiTrack Flex V100 cameras. The cameras are depicted as orange pyramids, the markers on the subject are colorized in blue, and the orange balls are outliers. The axes representation is x, y, z for red, green, and blue lines, respectively.

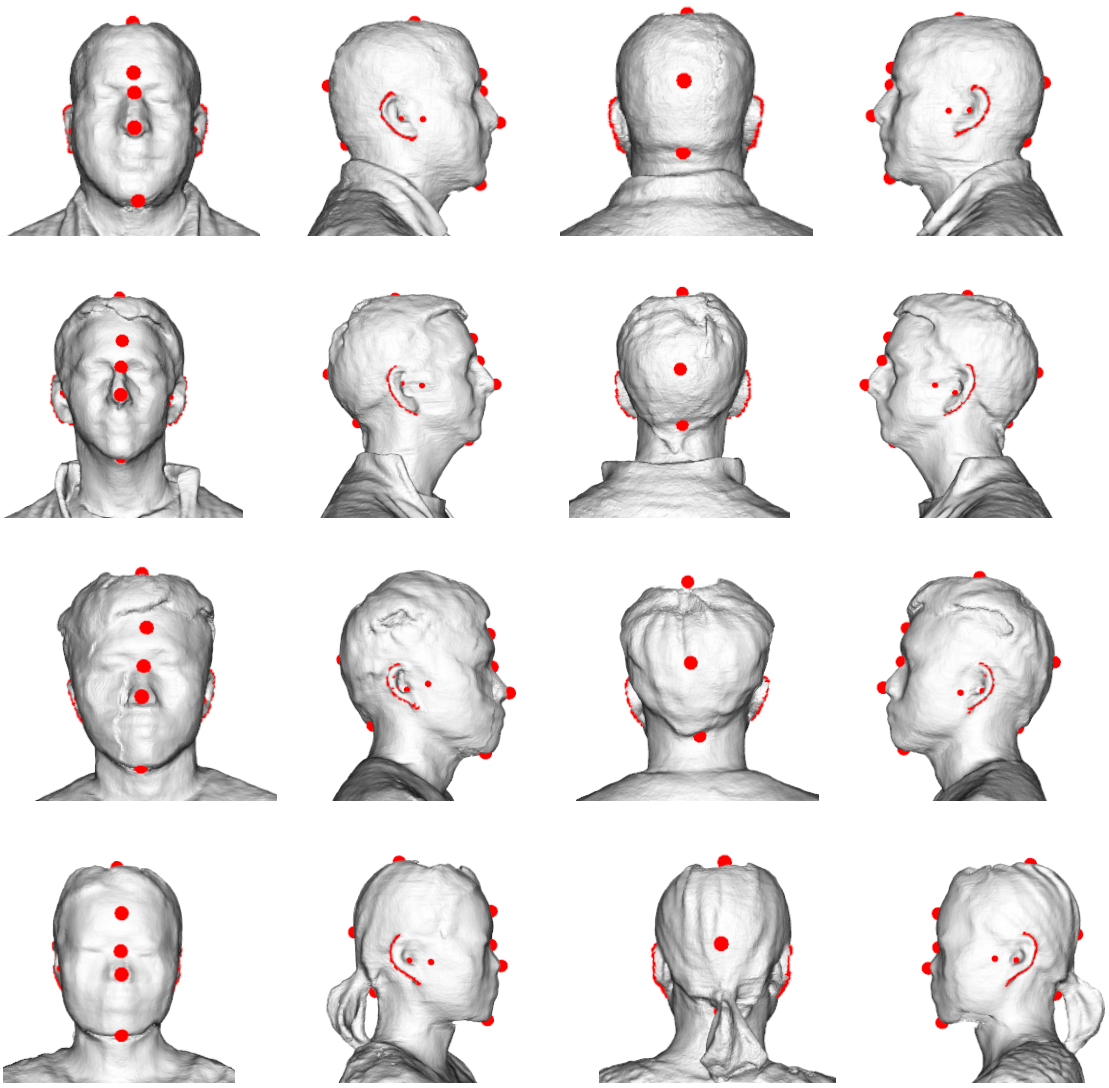


Figure 4.15: Visual results of feature points detection. The red circle represents each feature point. First and second rows: models taken with a Nvidia GeForce GT 630, rotation in clockwise and counter-clockwise direction, respectively. Third and fourth rows: models taken with a Nvidia GeForce GTX 560 Ti, rotation in clockwise and counter-clockwise direction, respectively.

Considering only the bad ear pit detections obtained in the presented dataset, the triangular fossa was the area in which the wrong ear pits were mostly localized (44.4 %), followed by the area behind the ear lobe (33.3 %) and behind the upper part of the helix (22.2 %). The estimation of the other feature points tended to not be affected to a large degree by a wrong ear pit detection (see Fig. 4.18) since the wrongly detected ear pit was, none the less, placed close to the concha. Furthermore, after analyzing the influ-

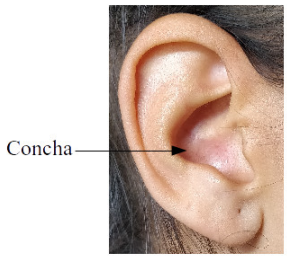


Figure 4.16: The ear pit detection was considered successful if the detected point was localized inside the volume of the concha.

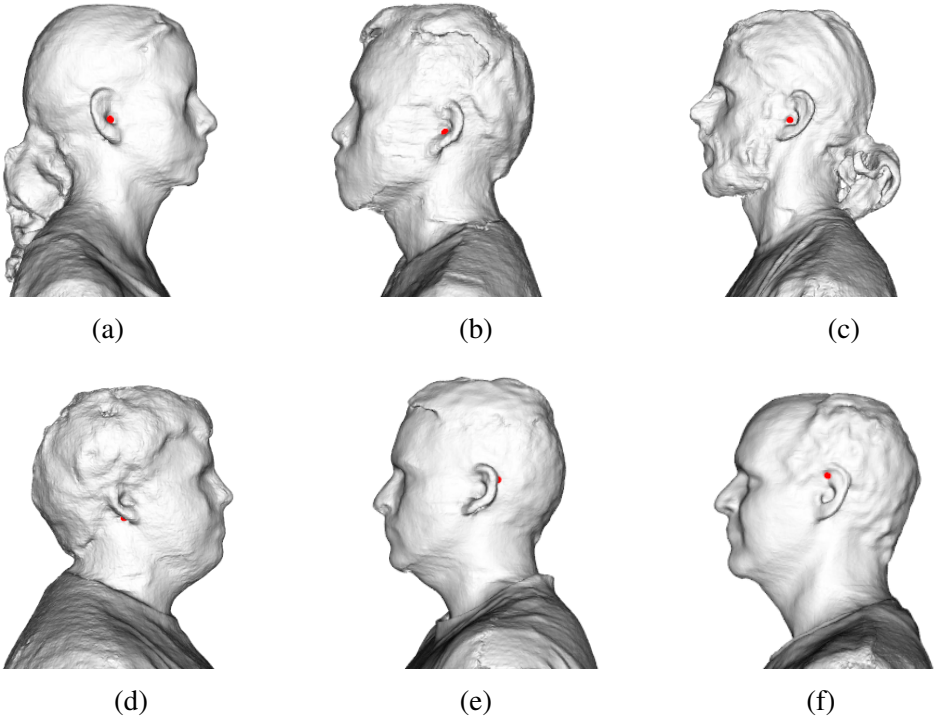


Figure 4.17: The ear pit is colorized in red. Top: ear pits correctly localized. Bottom: ear pits wrongly localized. (d) The ear pit was localized behind the lobule. (e) The ear pit was detected behind the upper part of the helix. (f) The ear pit is estimated on the triangular fossa.

ence of the user's turning direction on the badly detected ear pits, the results showed that among all the badly detected ear pits, 66.7 % were located on the ear that was seen last by the sensor, i.e. the left ear when turning in counter-clockwise direction, and right ear when turning clockwise. Nevertheless, the bad-ear-pit-located population is too small to establish a tendency.

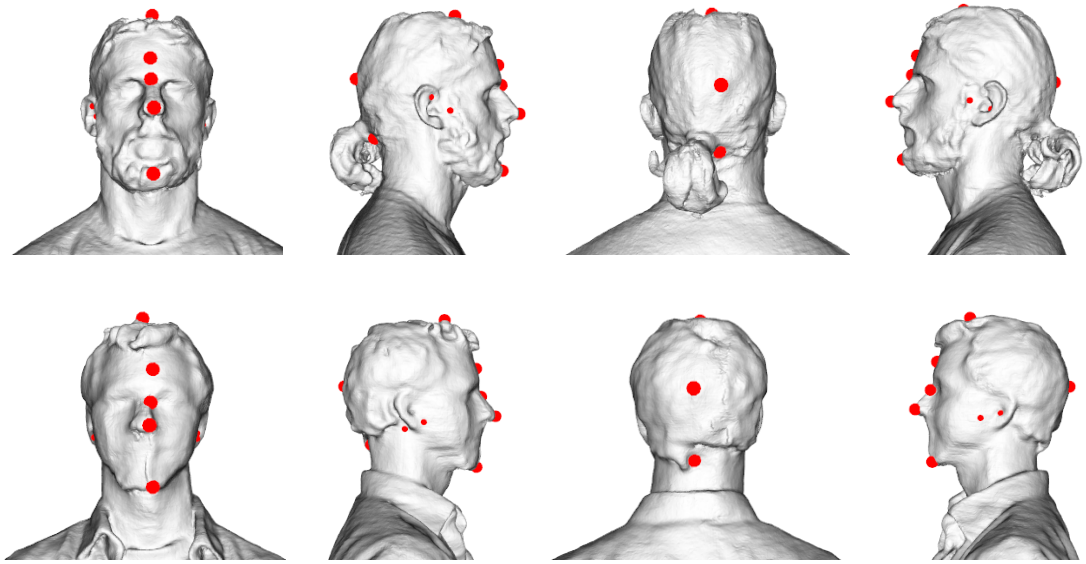


Figure 4.18: Visual results of feature points detection with wrongly localized ear point. Top: the ear pit of the right ear was located on the triangular fossa. Bottom: The ear pit of the right ear was located behind the ear lobe.

4.3.5 Accuracy evaluation

The accuracy evaluation was performed through the comparison of the estimated dimensions with their respective ground-truth values (the estimated head width was compared with the ground truth obtained with the tracking system, as well as with the caliper). The comparison consisted in solving the equation

$$e_m = m_e - m_t, \tag{4.2}$$

where e_m is the accuracy error, m_e is the measurement given by the algorithm, and m_t is the mean of the measurements[1] given by the tracking system or the caliper. Figure 4.19 shows the result of such comparison, all measurements but the helix height compared the 80 samples. Since the helix was occluded or not recognizable for some subjects, only 40 samples (80 ears) were selected for evaluation. However, two out of the 80 ears available for the helix detection presented artifacts and were removed from the evaluation, resulting in the analysis of 78 helix heights.

The absolute median error ranged between 2.1 mm and 21.5 mm, whereas the mean varied from 2.7 mm to 21.4 mm. The smallest mean and median values were given by the estimation of the helix height, the largest by the head depth. The head width and depth are mostly overestimated by the algorithm, whereas the height, nose bridge, and nape location, as well as the helix height do not present a tendency to be either over- or

[1] Each measurement was acquired three times.

underestimated. According to the signed accuracy error, five measurements presented outliers:

1. the width obtained with both the caliper and the tracking system, where the three outliers represent the same subject. These values stand out because the measurement obtained on the subject is closer to the estimated ones than in the case of other samples. Such similar measurements might be due to smaller horizontal displacements when the subject rotated in comparison to the other subjects,
2. the head depth, where the three outliers belong to the same subject. These values describe the posture change of the subject during the experiments,
3. the distance from the top to the nape, where the position of the nape's subject was erroneously estimated,
4. the helix height, where one of the outliers is given by a bad helix estimation. The other one is due to the detection of a section of the subject's beard as part of the helix on account of the beard's high curvatures.

Systematic error analysis

In order to understand the source of the accuracy errors, a systematic error analysis was performed. Figure 4.20 shows the ground-truth values compared to the measurements obtained with the algorithm. Offset errors were observed for the three measurements computed on the horizontal axes: width measured with both the caliper and the tracking system, and the depth. Such errors might be due to a bad sensor calibration and to the subjects' horizontal displacements when rotating. The offset values were computed as the average difference between the ground-truth data and the estimated values of all the subjects for each affected measurement.

The accuracy after the systematic error correction is shown in Fig. 4.21. The absolute median error varies between 2.1 mm and 13.3 mm, while the mean ranged from 2.7 mm to 12.2 mm. The smallest mean and median values are still given by the helix height estimation; however, the largest median and mean errors are now obtained when locating the nose bridge and the nape, respectively.

After the offset correction, the absolute median and mean errors of the width measured with the caliper were reduced by 13.2 mm and 11.4 mm, respectively. For the width measured with the tracking system the reduction was of 11.1 mm and 9.8 mm for median and mean, respectively. Finally, the depth's median and mean errors decreased 18.5 mm and 17.7 mm, respectively.

Comparison of the presented method to the state of the art

Other methods to estimate anthropometric dimensions from point clouds are not known by the author. However, a comparison of the proposed approach to the state-of-the-art

works of Torres-Gallegos et al. [97] and Dinakaran et al. [35] is presented in Table 4.2. Torres-Gallegos et al. developed an algorithm based on RGB images taken in a controlled environment, Dinakaran et al. use 3D models after manual post-processing. Meanwhile, the algorithm presented here required neither a controlled environment nor post-processing. The mean absolute error of the head width and depth given by the presented approach are smaller than the one reported by Torres-Gallegos et al. but larger than the result given by Dinakaran et al. Regarding the head and helix height, the presented algorithm exhibits the largest errors: 2.6 mm and 0.3 mm more than the highest value of the compared algorithms for head and helix height, respectively.

	Mean absolute error		
Measurement	Torres-Gallegos' [mm]	Dinakaran's [mm]	Presented approach's [mm]
Head width	6.3	2.2	4.3 & 4.1*
Head depth	4.1	1.8	3.7
Head height	6.5	3.6	9.1
Helix height	0.7	2.4	2.7

Table 4.2: Comparison of the accuracy of the presented approach with the state of the art. *The ground truth related to the first value was obtained with the caliper, the ground truth of the second value was obtained with the tracking system.

4.3.6 Reproducibility evaluation

The reproducibility was evaluated as the standard deviation of the estimated measurements of each subject's four models (which were taken with the two different graphic cards and the two rotation directions). For the helix standard deviation computation, the samples with artifacts were removed.

The median and mean standard deviation, shown in Fig. 4.22, varied from 1.9 mm to 4.1 mm and from 2.2 mm to 5.0 mm, respectively. Since the results were equal or lower than 5 mm, they indicated that the presented approach can be used with different machines and without restrictions of turning direction. However, in order to analyze if better results were obtained in specific conditions, Table 4.3 and Table 4.4 show the accuracy error classified by both the user's rotation direction and the graphic card used during the reconstruction (for the helix height evaluation, the samples with artifacts were removed). Smaller errors were exhibited when the user turned clockwise, which might expose a major turning control by the subject when rotating to the right. Regarding the helix height, a correlation between the analyzed ear and the turning direction was exhibited. When the user turned clockwise the right helix height presented smaller errors than the left helix height; likewise, when the rotation was counter-clockwise, the left helix height was estimated more accurately than the right one. Additionally, a relation

between the graphic card used and the accuracy was not detected. Furthermore, an association among the rotation direction, the graphic card, and the measurements over- or underestimation was not found.

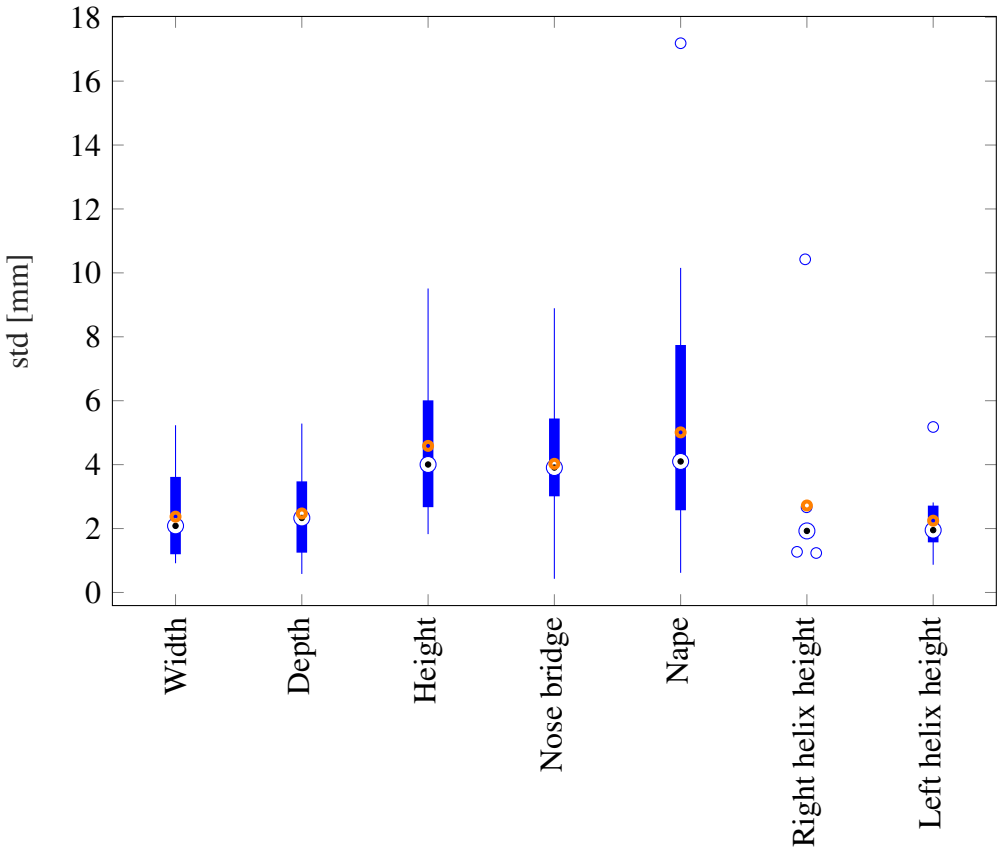


Figure 4.22: Reproducibility evaluation. The topmost and bottommost limits of the blue lines represent the maximum and minimum values, respectively. The upper and lower lines of the blue boxes describe the third and first quartile, respectively. The black dots give the median, and the orange rings the mean. Outliers are symbolized as blue circles.

4.4 Summary and Conclusions

In this chapter, an approach to automatically estimate head and ear measurements from uncolored 3D point clouds is presented and evaluated. The point clouds were generated using an out-of-the-shelf RGBD sensor called Xtion Pro Live fabricated by Asus and the software KinFu.

The estimated head and ear measurements were defined by the requirements to personalize head related transfer functions, and they are the head width, depth and height,

Measurement	Nvidia GeForce GT 630						Nvidia GeForce GTX 560 Ti					
	Clockwise			Counter-clockwise			Clockwise			Counter-clockwise		
	median	μ	std	median	μ	std	median	μ	std	median	μ	std
Head width*	**15.8**	15.7	4.8	16.7	16.4	4.4	**15.8**	**15.0**	4.6	16.9	15.6	5.4
	14.4	14.0	5.1	15.1	14.5	5.2	14.2	**13.1**	4.6	**14.0**	14.1	5.8
Head depth	**19.7**	**20.3**	4.6	22.8	22.5	5.3	21.1	20.4	5.6	23.0	22.3	4.7
Head height	8.8	**7.9**	5.9	**8.5**	9.4	6.6	10.0	9.4	6.3	8.6	9.5	6.7
Nose bridge	**10.9**	**10.2**	6.5	13.8	11.8	6.1	14.9	13.4	6.4	13.5	12.2	6.7
Nape	11.7	13.9	10.0	8.6	**11.3**	8.9	**8.3**	**11.3**	9.3	8.7	12.3	12.4
Helix height (R)	2.9	2.8	1.7	3.0	4.7	6.2	**1.5**	**2.3**	2.2	2.3	2.6	1.6
Helix height (L)	2.1	2.7	1.9	2.3	2.6	1.8	**1.9**	1.9	1.2	**1.9**	**1.7**	1.0

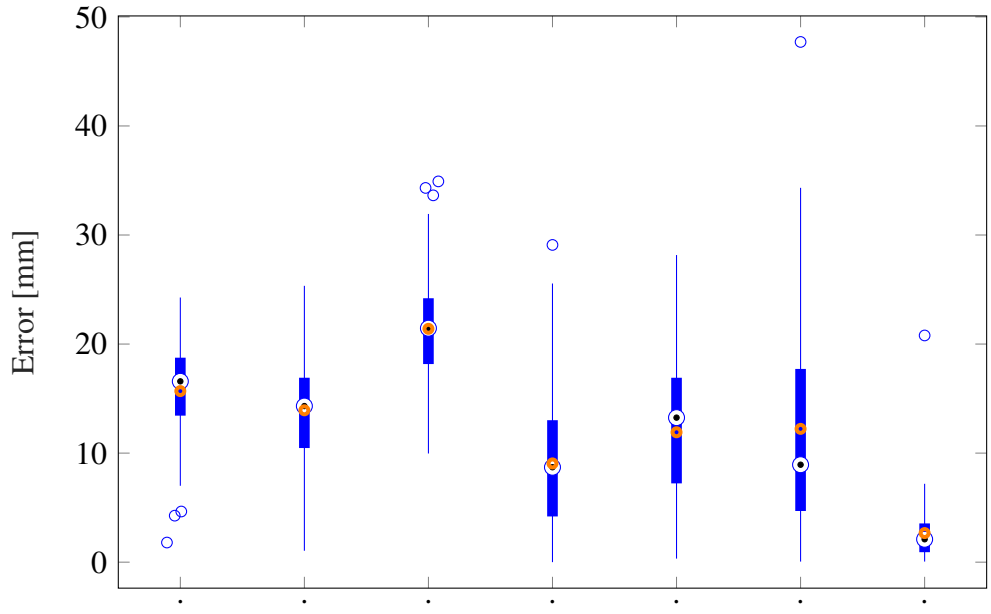
Table 4.3: Absolute error (in millimeters) classified by graphic card and rotation direction. The highlighted numbers represent the lowest median and mean value obtained per measurement. *The ground truth related to the first-line values was obtained with the caliper, the ground truth of the second-line values was obtained with the tracking system.

Measurement	Nvidia GeForce GT 630						Nvidia GeForce GTX 560 Ti					
	Clockwise			Counter-clockwise			Clockwise			Counter-clockwise		
	median	μ	std	median	μ	std	median	μ	std	median	μ	std
Head width*	**15.8**	15.7	4.8	16.7	16.4	4.4	**15.8**	**15.0**	4.6	16.9	15.6	5.4
	14.4	13.8	5.5	15.1	14.5	5.2	14.2	**13.0**	4.9	**14.0**	13.7	6.7
Head depth	**19.7**	**20.3**	4.6	22.8	22.5	5.3	21.1	20.4	5.6	23.0	22.3	4.7
Head height	**-0.4**	**0.2**	10.1	4.1	2.5	11.4	4.4	2.9	11.1	4.5	3.8	11.1
Nose bridge	**3.4**	**3.8**	11.7	9.0	4.8	12.6	12.3	6.9	13.4	9.5	6.3	12.7
Nape	-11.7	-10.9	13.4	-8.0	**-8.2**	12.0	-7.0	-8.9	11.7	**-5.8**	-9.5	14.8
Helix height (R)	-1.5	-0.8	3.3	-1.5	-3.8	6.8	**-0.2**	**-0.4**	3.3	-1.6	-1.2	2.9
Helix height (L)	-2.0	-1.9	2.8	-1.7	-0.7	3.2	-0.9	-0.8	2.2	**-0.7**	**-0.5**	2.0

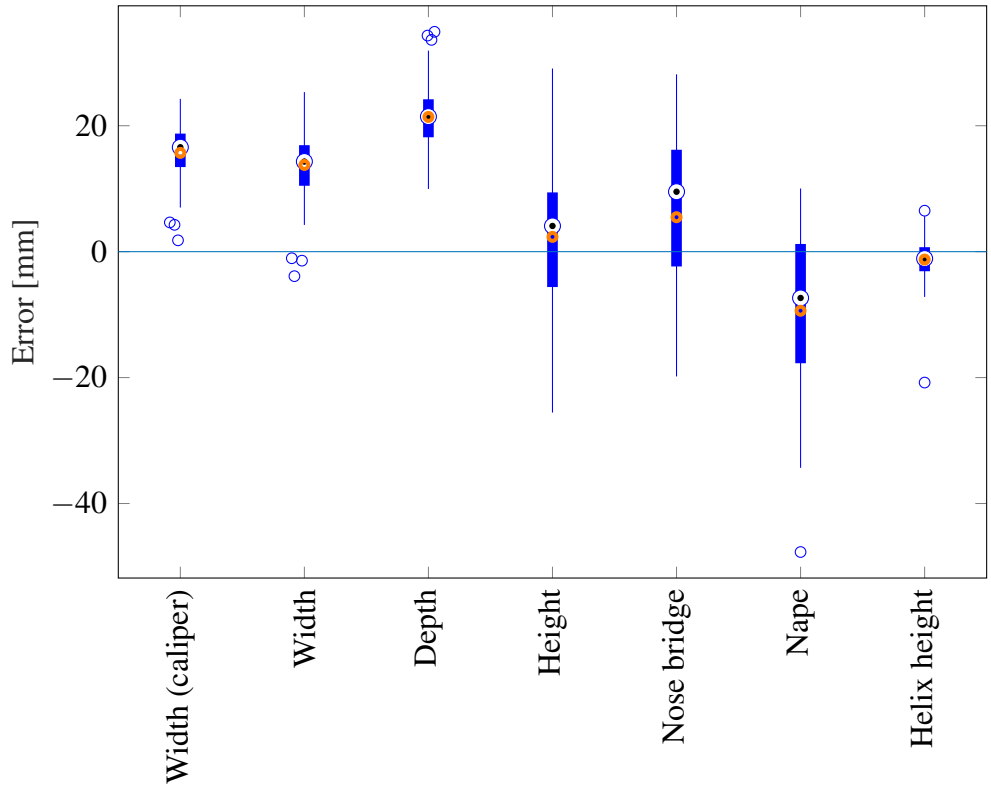
Table 4.4: Signed error (in millimeters) classified by graphic card and rotation direction. The highlighted numbers represent the closest to zero median and mean value obtained per measurement. *The ground truth related to the first-line values was obtained with the caliper, the ground truth of the second-line values was obtained with the tracking system.

the distance from the nose bridge and the nape to the ear pits, as well as the helix height. In order to estimate such measurements, feature points were localized on the point cloud of the head, e.g., ear pits, nose bridge, and nape, among others. For the evaluation of the algorithm, point clouds sets (i.e. each set contains one 360° and two profile -left and right- point clouds) of 20 subjects were acquired at two different frame rates when each subject rotated in clockwise and counter-clockwise direction. Therefore, the head measurements were evaluated using 80 point clouds sets. Nonetheless, since the helix was not clearly delimited on some reconstructions, the evaluation of the helix height estimation was performed using the point clouds sets of 10 subjects (40 point clouds sets). Experimental results showed an ear pit localization accuracy of 94.3 %. An analysis of the measurements' accuracy showed the presence of systematic errors (offset) in the measurements measured along the horizontal axes, such as head width (both compared to the caliper and the tracking system ground truths) and depth. After the systematic error correction, the range of the median accuracy error was between 2.1 mm to 13.3 mm, while the mean ranged from 2.7 mm to 12.2 mm. The estimation of the helix height exhibited the smallest mean and median values; the largest median and mean errors were given by the estimation of the distance to the nose bridge and the nape, respectively. In order to evaluate the reproducibility of the algorithm, a comparison among the four sets of each subject was performed through the computation of the standard deviation. The results showed a median and a mean standard deviation that varied from 1.9 mm to 4.1 mm and from 2.2 mm to 5.0 mm, respectively. Since the measurements deviation among the 3D models of each subject was small, the algorithm is categorized as reproducible on data acquired under different conditions.

In order to decrease the systematic error, a correct camera calibration as well as a proper indication to the user about the rotation procedure (to avoid displacements while rotating) are recommended. Furthermore, in the present study a relation between the graphic card used, which indicates the 3D resolution and the reconstruction rate, and the measurement accuracy error was not found.

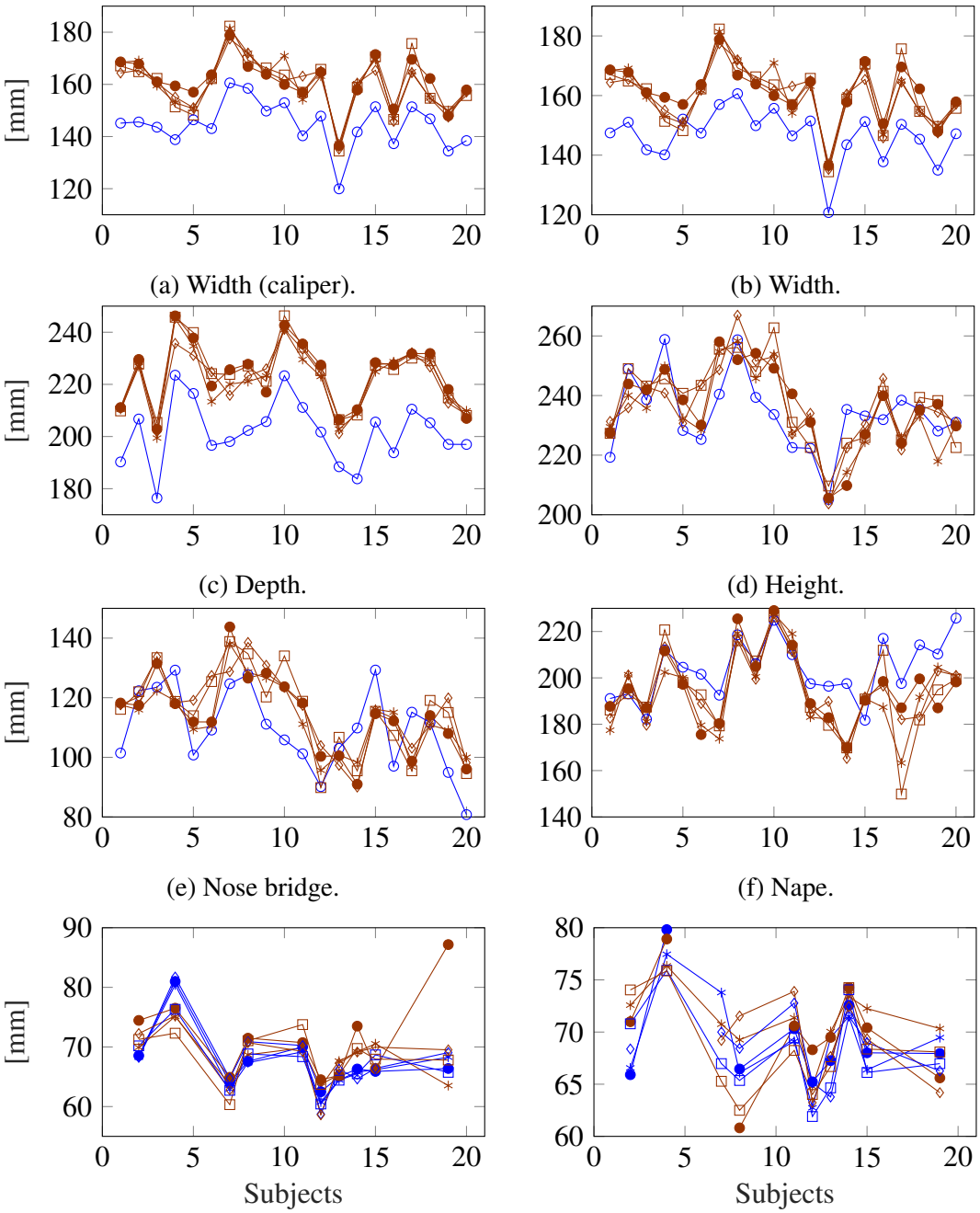


(a) Absolute accuracy error.



(b) Signed accuracy error

Figure 4.19: Accuracy evaluation. The topmost and bottommost limits of the blue lines represent the maximum and minimum values, respectively. The upper and lower lines of the blue boxes describe the third and first quartile, respectively. The black dots give the median, and the orange rings the mean. Outliers are symbolized as blue circles.



(a) Width (caliper).

(b) Width.

(c) Depth.

(d) Height.

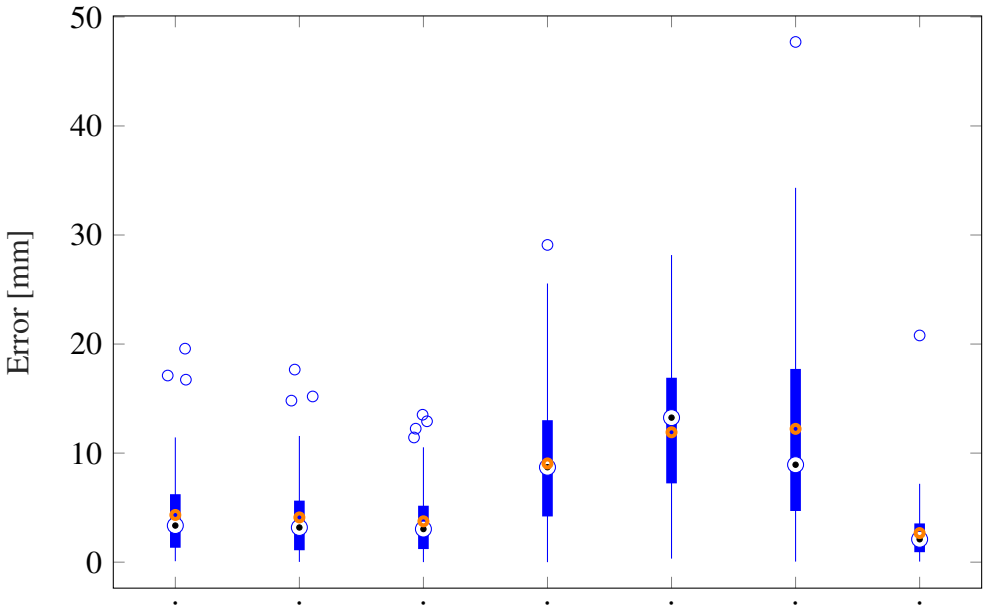
(e) Nose bridge.

(f) Nape.

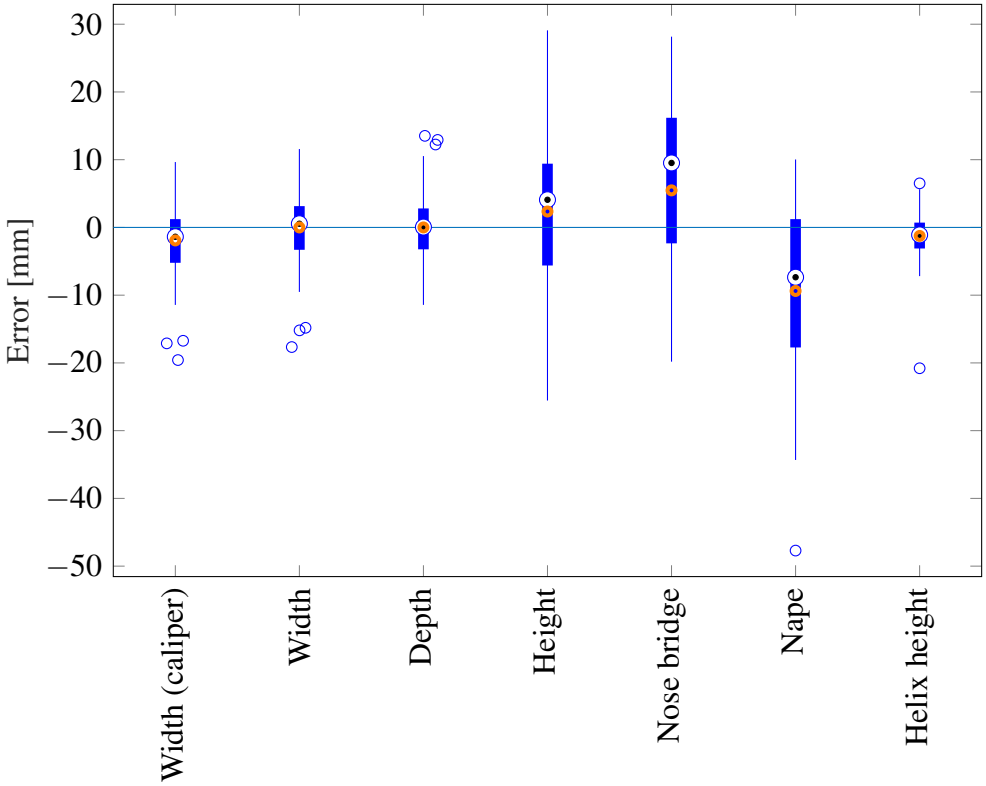
(g) Right helix height.

(h) Left helix height.

Figure 4.20: Comparison between the ground-truth values and the estimated measurements. The blue markers represent the ground truth, the brown markers symbolize the measurements given by the algorithm. ○ describes the ground-truth measurements taken on the subjects. ∗ and • are from models acquired at ≈ 5 fps while the subject turned in clockwise and counter-clockwise direction, respectively, ◇ and □ are from models acquired at ≈ 10 fps while the subject turned in clockwise and counter-clockwise direction, respectively.



(a) Absolute accuracy error.



(b) Signed accuracy error.

Figure 4.21: Accuracy evaluation after systematic error correction. The topmost and bottommost limits of the blue lines represent the maximum and minimum values, respectively. The upper and lower lines of the blue boxes describe the third and first quartile, respectively. The black dots give the median, and the orange rings the mean. Outliers are symbolized as blue circles.

Chapter 5

A Case Study of Head Measurements in Acoustics: Influence of the anthropometric measurements on the interaural time difference

Humans localize sound sources using the Head Related Transfer Function (HRTF), which relies on person-dependent characteristics such as head, torso, external and internal ear dimensions, among others. The capability of locating sound sources improves the immersion of people in their environment; unfortunately, when headphones are used, humans cannot locate sound sources in a 3D space unless such headphones contain the corresponding HRTF. Obtaining each person's HRTF is a tedious task since it requires specialists and a dedicated laboratory. Therefore, in order to find more practical methods to personalize this function, several approaches exist to either select the HRTF from an existing database [75, 104, 63, 50] or to estimate it using anthropometric data [20, 97, 35]. The presented case study is based on the latter kind of approach.

The most important sound source localization cues can be classified as interaural time difference (-ITD- difference between the time that the sound takes to reach each ear [89]) at lower frequencies and interaural level difference (difference in loudness and frequency between the sound that reaches each ear [89]) and spectral cues at higher frequencies [81, 62], they are both related to the anthropometric information of each person. Hence, the objective of the presented case study is to validate the automatic head measurement estimation approach described in Chapter 4 at low frequencies, i.e., when computing personalized ITDs.

This chapter is organized as follows. The theoretical input accuracy for two ITD models is estimated in Section 5.1. The description of the experiment to validate the feasibility of the developed algorithm (in Chapter 4) and its evaluation are presented in Section 5.2 and Section 5.3, respectively. The influence of the estimated head measurements on the calculation of ITDs is described in Section 5.4. Finally, Section 5.5 presents the summary and conclusions.

The presented chapter is based on the paper: *Required Measurement Accuracy of Head*

Dimensions for Modeling the Interaural Time Difference authored by Bomhardt, Patiño, Zell, and Fels [26]. The first and second authors contributed equally to the publication. Some fragments presented here are replicated from the original paper.

5.1 Input accuracy for ITD Models

The content of this section was developed by Ramona Bomhardt at the RWTH Aachen University, see [26].

A theoretical analysis was performed in order to estimate the required input accuracy of Woodworth's [103] and Kuhn's [54] ITD models, which are well-known models. Woodworth's model is used in the range of 0.6 KHz to 3.1 KHz, below this frequency range Kuhn's is more accurate. Such models are based on the head radius:

$$\mathrm{ITD}_{\mathrm{Wood}} = \frac{a}{c_0}(\sin\varphi + \varphi) \quad \text{for} \quad 0 \leq \varphi \leq \frac{\pi}{2}\,, \tag{5.1}$$

$$\mathrm{ITD}_{\mathrm{Kuhn}} = \frac{3a}{c_0}\sin\varphi \quad \text{for} \quad (ka)^2 \ll 1\,, \tag{5.2}$$

where φ is the azimuth angle on the horizontal plane, a describes the averaged head radius, the speed of sound is represented by c_0, and k is the wave number. The averaged head radius can be expressed as a function of the head width w_c, the head depth d_c and height h_c (see Fig. 5.1) using the Algazi approximation [20]:

$$a_{\mathrm{Algazi}} = 0.51w_c + 0.18d_c + 0.019h_c + 32\,\mathrm{mm}\,. \tag{5.3}$$

The just noticeable anthropometric measurement error $\Delta\mathrm{ITD}_{\mathrm{Wood}}$ and $\Delta\mathrm{ITD}_{\mathrm{Kuhn}}$ are determined using the propagation of uncertainty:

$$\Delta\mathrm{ITD}_{\mathrm{Wood}} = \frac{\partial\mathrm{ITD}_{\mathrm{Wood}}}{\partial a}\Delta a = \frac{\Delta a}{c_0}(\sin\varphi + \varphi)\ \text{, and} \tag{5.4}$$

$$\Delta\mathrm{ITD}_{\mathrm{Kuhn}} = \frac{\partial\mathrm{ITD}_{\mathrm{Kuhn}}}{\partial a}\Delta a = \frac{3\Delta a}{c_0}\sin\varphi\,, \tag{5.5}$$

both can be expressed in terms of the just noticeable difference (JND) of the ITD deviation (it should be previously calculated) assuming that the just noticeable measurement error is equal to the JND of the ITD deviation:

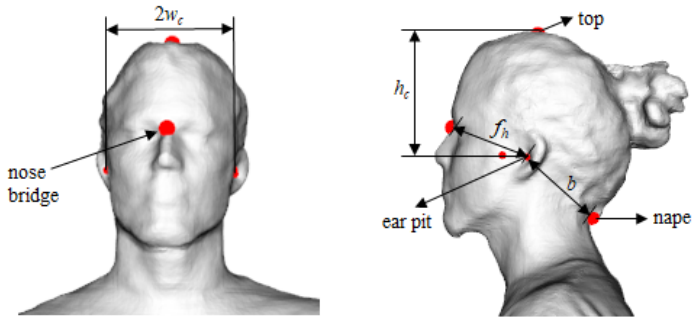


Figure 5.1: The head dimensions are labeled as follows: w_c is the head width divided by two, h_c represents the head height, the shortest distance from the ears to the front is f_h, the shortest distance from the ears to the back is b, and the depth $d_c = 0.5\,(f_h + b)$.

$$\Delta a_{\text{Wood}} = \frac{c_0}{\sin\varphi + \varphi} \cdot \text{ITD}_{JND}\ \text{, and} \tag{5.6}$$

$$\Delta a_{\text{Kuhn}} = \frac{c_0}{3\sin\varphi} \cdot \text{ITD}_{JND}\ . \tag{5.7}$$

The results showed that the minimum allowed measurement error is 7 mm when using Woodworth's method (Δa_{Wood}) and 5 mm for Kuhn's (Δa_{Kuhn}). Such values are considered as the maximum allowed measurement error for the head dimensions.

Using the head radius given by Algazy (a_{Algazi}) and ITD approaches such as eq. 5.1 and eq. 5.2, the input measurement uncertainty is

$$\begin{aligned}\Delta\text{ITD}_{\text{Algazi}} = {} & \frac{\partial \text{ITD}_{model}}{\partial w_c}\Delta w_c + \frac{\partial \text{ITD}_{model}}{\partial d_c}\Delta d_c \\ & + \frac{\partial \text{ITD}_{model}}{\partial h_c}\Delta h_c\ .\end{aligned} \tag{5.8}$$

5.2 Experiment

In order to validate the approach described in Chapter 4 when personalizing ITDs, 3D models of the upper body of 17 people were generated with an RGBD sensor Asus Xtion Pro Live and the software KinFu. The subjects' age was 29 $\pm$ 5 years old, 13 were male and 4 females. The HRTFs and anthropometric information of these subjects are available in the ITA HRTF database [25].

The estimated anthropometric dimensions are shown in Fig. 5.1 and are defined as

follows:

- w_c is the head width, which is the distance measured in front of the tragi from left to right divided by two,
- f_h represents the shortest distance from the ear pits to the front of the head,
- b is the shortest distance from the ear pits to the nape,
- d_c symbolizes the head depth calculated as $d_c = 0.5\,(f_h + b)$,
- h_c is the head height, representing the distance from the ear pits to the top of the head.

The method explained in Section 4.2 allows the direct computation of all the aforementioned measurements but the head height, which is calculated (on the plane yz) as the distance from the top of the head to the mean position of the ear pits.

The process to acquire the 3D models in the current case study was similar to the procedure described in Section 4.3.1. However, the RGBD sensor Asus Xtion Pro Live was placed at 0.74 m (mean distance among all the subjects) from the subject and used to record the 3D point clouds of the upper body. Such point clouds were later processed offline with the KinFu application available in PCL. The 3D reconstructions were generated using a computer with a graphic card Nvidia GeForce GTX 970. The resolution for the reconstructions was 512 voxels per axis.

5.3 Evaluation of the Estimated Anthropometric Data

5.3.1 Qualitative evaluation

In the 3D models, the ear pit was detected as a point on the concha. Therefore, three subjects were not taken into account for the evaluations, since the ear pit was detected outside the ear region on at least one ear. This as result of lack of concavity on the ear pit or/and bad alignment of the subject with respect to the sensor. Additionally, the ear pit of one of the subjects was localized on the triangular fossa. Nonetheless, since the triangular fossa is notable close to the concha, the detected feature points required for the evaluation of the ITD were not influenced by the bad localization of the ear pit. Thus, the subject was involved in the measurements and ITD evaluations.

The 3D models generated using KinFu represent the subject's current physical appearance. Consequently, the hair style is part of the model and may affect the automatic estimation of the head dimensions, particularly the head height as is shown in Fig. 5.2. The top of the head is wrongly estimated on the pompadour of one of the subjects, whereas the other subject presented a correct top estimation due to his lack of hair. Given that among the participants in this study only two subjects were bald, the calculation of

twelve subjects' head height was considered influenced by the hair. Additionally, after a visual evaluation of the head features location, it was inferred that the hair drastically modified the nape position estimation of three subjects as is shown in Fig. 5.3. Therefore, the distance to the nape of these subjects was also considered inaccurate. Due to the influence of the hair on the calculation of h_c and b, the anthropometric data of the affected subjects were estimated from existing anthropometric databases by an alternative approach explained in Section 5.3.2.

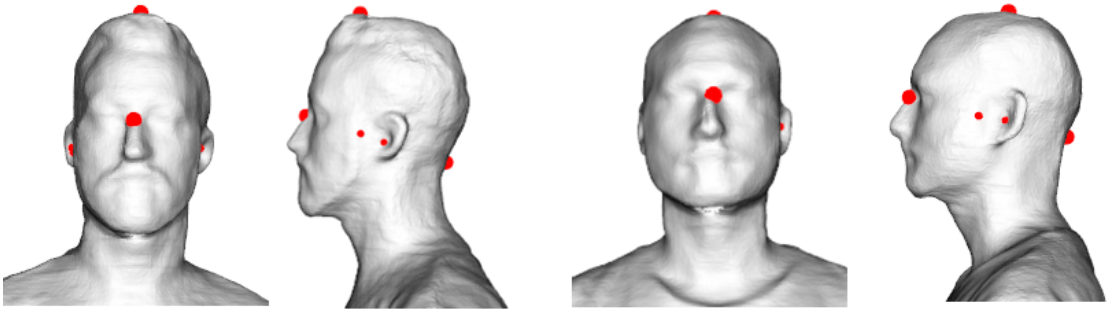


Figure 5.2: Visual results of head features detection for two different subjects. For both, the frontal view and profile are depicted. The detected features are marked by red dots.

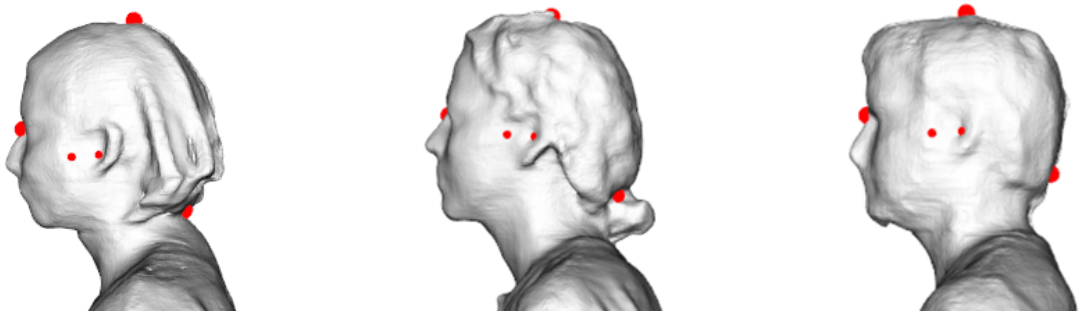


Figure 5.3: The hair affected the nape position estimation for these three subjects.

5.3.2 Alternative approach for the estimation of the head height and depth

As explain in Section 5.3.1, the hair style at the top of the head and at the nape influenced the automatic detection of the required feature points to calculate the head height h_c and the shortest distance from the ear pits to the back b for some subjects (see Fig. 5.2 and Fig. 5.3). Therefore, these dimensions are estimated after merging two anthropometric databases: CIPIC and ITA HRTF database. Since the 17 subjects of the current case

study are also considered in the ITA HRTF database, they were removed from the merged database.

Contrary to the information in the ITA HRTF database, the definition of the CIPIC's head height differs from the one required in this study. However, it can be derived using the CIPIC's head height d_{x_2} (measured from the top to the chin) and the vertical pinna offset d_{x_4}

$$h_c = \frac{d_{x_2}}{2} + d_{x_4} \,. \tag{5.9}$$

The shortest distance from the ear pits to the back b is computed from the CIPIC dimensions using the CIPIC's head height d_{x_2}, the CIPIC's head depth d_{x_3} (measured from the forehead to the back of the head), the vertical pinna offset d_{x_4}, and the horizontal pinna offset d_{x_5}

$$b = \sqrt{\left(\frac{d_{x_3}}{2} - d_{x_5}\right)^2 + \left(\frac{d_{x_2}}{2} - d_{x_4}\right)^2} \,. \tag{5.10}$$

In order to estimate the hair-influenced head dimensions, the automatic estimated head width $2w_c$ of the investigated subject[1] is used to find, in the merged database, other subjects with the same gender whose head widths are in a comparable range ($\pm$ 5 mm). Based on the found subjects, an averaged anthropometric dimension was calculated.

In case this alternative approach was also unsuccessful[2] (i.e. subjects with a similar head width were not found in the merged database), the missing anthropometric dimension was calculated as the mean of the required measurement of all the subjects in the merged database.

5.3.3 Accuracy evaluation

In order to evaluate the automatically-acquired anthropometric data accuracy, a comparison to ground-truth values was performed. The ground truth was obtained from magnetic resonance image (MRI) scans. Such scans were taken from March 2015 to June 2016 at the "Radiologische, Nuklearmedizinische und Strahlentherapeutische Gemeinschaftspraxis" in Aachen, Germany. They are published in an open-access database by Bomhardt et al. (ITA database). Figure 5.4 shows the localization of the feature points required to estimate the head measurements on a set of MRIs. The head width w_c is measured between the head center and the region in front of the tragus, whereas the distance from the center to the nose bridge and to the nape represent f_h and b, respectively. The

[1] The approach is based on the fact that the head dimensions are correlated. Considering the head dimensions of the 48 subjects in the ITA HRTF database, the correlation coefficient of the head width w_c and depth d_c is $\rho_{wd} = 0.6$. The correlation coefficient of the head width w_c and height h_c is slightly lower $\rho_{wh} = 0.4$.

[2] This only happened to one subject in the presented study.

head height h_c was measured between the head center and the top.

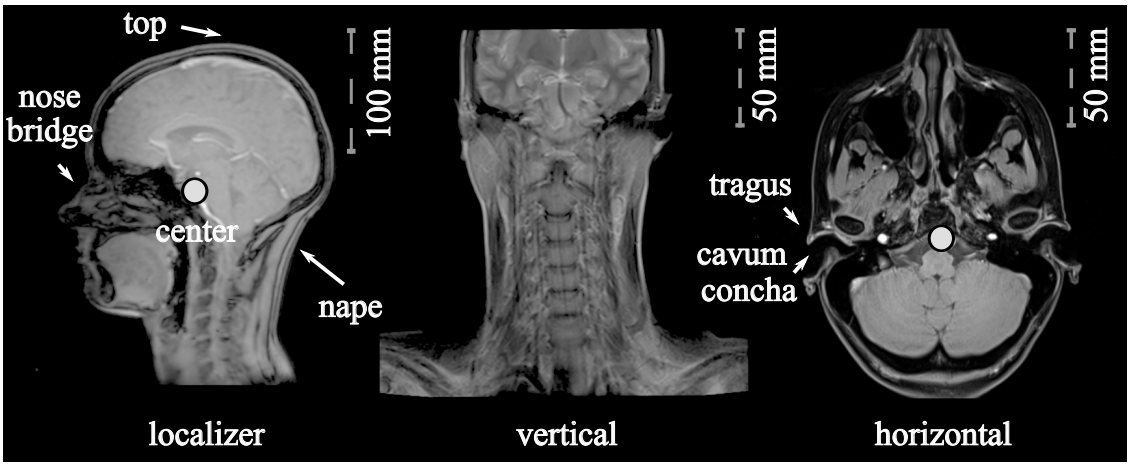


Figure 5.4: Feature points to obtain ground-truth measurements. A point in front of the tragus to the center is used to calculate the head width w_c, f_h and b are measured from the center to the nose bridge and the nape, and the head height h_c is the distance from the top to the center.

The measurements accuracy was evaluated on 14 subjects (three subjects were excluded due to bad ear pit localization, see Section 5.3.1) following the equation

$$e_c = m_t - m_e, \tag{5.11}$$

where e_c is the accuracy error, m_t is the ground-truth value, and m_e symbolizes the estimated measurement obtained from the 3D head model.

The results presented in Fig. 5.5 shows that the approaches used in the presented case study to automatically estimate the head measurements tend to overestimate the dimensions. Additionally, the distance between the ear pit and the nape presents a standard deviation larger than the other measurements. The reason behind this behavior might be the influence of the hair style on the nape localization.

The absolute mean error of the head width ($2w_c$) is 6 mm (here the width is expressed as the distance from right to left of the head for comparison purposes), it is comparable to the one exhibited by Torres-Gallegos [97]. However, the error presented by Dinakaran [35] is smaller than the one presented in the current study. The other distances are not compared with the state of the art, since they are not measured by the other authors.

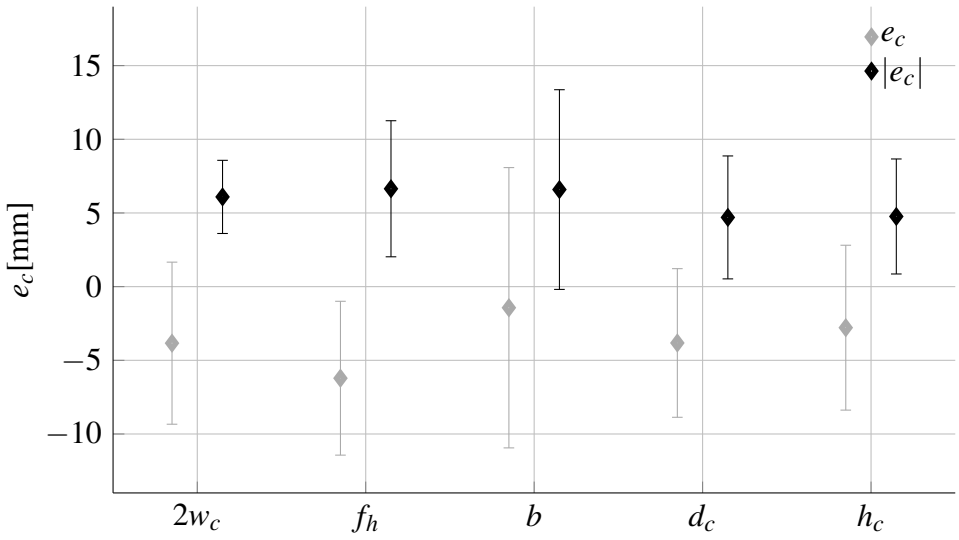


Figure 5.5: Accuracy evaluation of the anthropometric dimensions. The diamonds mark the mean error of the remaining 14 subjects and the bar its standard deviation. The black bars represent the unsigned errors $|e_c|$.

5.4 Effects of the Input Data Error on the Interaural Time Difference Calculation

The content of this section was developed by Ramona Bomhardt at the RWTH Aachen University, see [26].

The ITD error can be calculated using the differences between the ground-truth and the estimated anthropometric dimensions. Therefore, the mean absolute error of the head width (w_c), the head depth (d_c), and the head height (h_c) was replaced in eq. (5.8). Fig. 5.6 shows the resulting error $\Delta\text{ITD}_{Est.Val.}$, which compared to ΔITD_{JND} presents a smaller average error and standard deviation for all the tested directions. Thus, only very sensitive listeners will perceive a very small shift of the presented sound source position. These results validate the approaches presented in Sections 4.2 and 5.3.2 as suitable methods to personalize ITDs.

5.5 Summary and Conclusions

In order to prove the feasibility of the algorithm explained in Chapter 4, the automatically obtained head measurements were used to calculate the interaural time difference (ITD) of 14 subjects. First, an evaluation of the measurements accuracy was performed taking as ground-truth values the measurements obtained from magnetic resonance images. The results reflected that the automatically estimated dimensions tended to be overestimated. However, the accuracy mean error (which varied from 3.0 mm to 5.0 mm) was below the

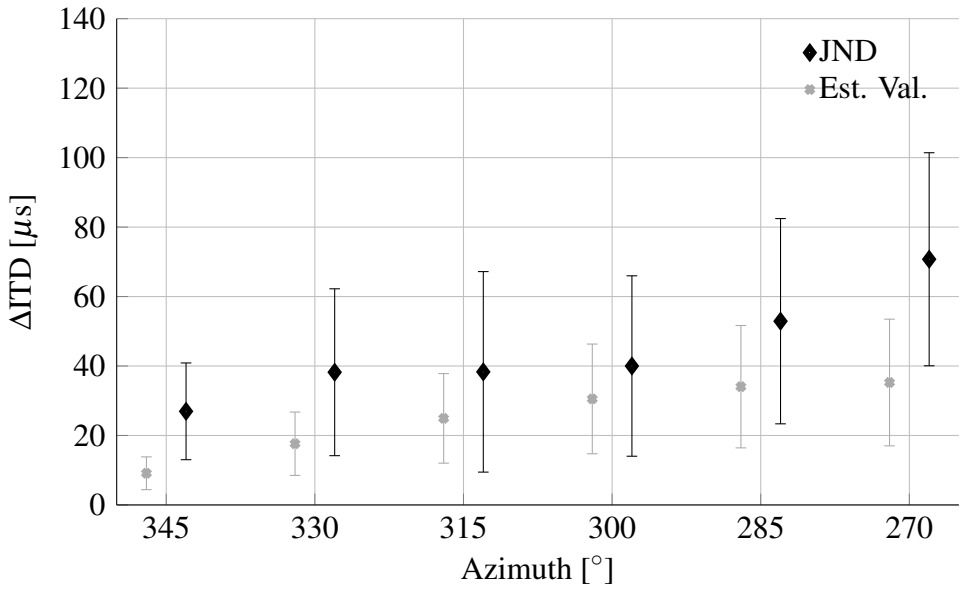


Figure 5.6: The JND of the ITD deviation ΔITD as well as the ITD error calculated with the automatically estimated anthropometric data $\Delta ITD_{Est.Val.}$.

established theoretical threshold for the calculation of ITDs. Second, the ITDs were calculated using the estimated measurements and it was concluded that only very sensitive listeners will perceive a small shift on the location of the sound source.

Regarding this acoustic application, it is proved that the algorithm to automatically estimate head measurements is an accurate and faster alternative to manual approaches.

Chapter 6

Accuracy of RGBD Sensors

As seen in Chapter 4, the combination of an RGBD sensor with algorithms such as KinFu proved to be a great alternative to generate accurate 3D reconstructions. Moreover, since the launch of the Microsoft Kinect in 2010, multiple different out-of-the-shelf RGBD sensors have been acquired by researchers and end users. These sensors employ structured-light, time-of-flight, and stereoscopic technology; a detailed explanation about such sensors is given in Chapter 2. Due to the amount of available RGBD sensors, the selection of one to generate 3D reconstructions can be a tedious task. For this reason, in this chapter an analysis of 3D reconstructions produced by KinFu is presented. The reconstructions were generated at different reconstruction rates and with different amount of input point clouds, when using various RGBD sensors. This with the aim of evaluating multiple scenarios that researchers and end users might find: 1. low and full reconstruction rate, possible due to the graphic card, 2. various distances between the sensor and the object to reconstruct, 3. uneven amount of input point clouds as in [29, 71], where the amount of input point clouds is not specified, and 4. low number of input point clouds[1]. Four objects were reconstructed: a dummy head, a soccer ball, an American football, and a rubber duck, see Fig. 6.1. The 3D information of the aforementioned objects was acquired while each object rotated, at different distances, in front of six sensors (see Fig. 6.2): a Microsoft Kinect Xbox 360, a Microsoft Kinect for Windows in near mode, an Orbbec Astra S, an Asus Xtion Pro Live, a Microsoft Kinect v2, and an Intel RealSense R200. The analysis includes qualitative and quantitative evaluations, namely: 1. success rate, 2. quality, 3. accuracy per point, 4. curvature accuracy, and 5. number of points of the reconstructions. For the quantitative evaluations, the 3D representation, as point cloud, of each object acquired with KinFu was compared to the respective ground-truth mesh. Such ground truth was obtained with an Artec Eva 3D Scanner, whose 3D point accuracy is up to 0.1 mm [1].

The remainder of this chapter paper is organized as follows. Section 6.1 presents the related work regarding the use of KinectFusion. A condensed explanation of the software and sensors used in the experiments is given in Section 6.2. The data acquisition process and the evaluation of the reconstructions are described in Section 6.3 and Section 6.4, respectively. Finally, Section 6.5 presents the summary and conclusions.

[1]The point clouds should be overlapped to work with KinFu.

Figure 6.1: Objects reconstructed using KinFu and different RGBD sensors. From left to right: dummy head, soccer ball, American football, and rubber duck.

Figure 6.2: Sensors used in the experiments. From top left to bottom right: Microsoft Kinect Xbox 360, Microsoft Kinect in near mode, Microsoft Kinect v2, Orbbec Astra S, Intel RealSense R200, and Asus Xtion Pro Live.

This chapter is based on the paper: *3D Reconstructions with KinFu Using Different RGBD Sensors* authored by Patiño and Zell [77]. Some fragments presented here are replicated from the original paper.

6.1 Related Work

The possibility to use KinectFusion as an algorithm for ground truth acquisition was studied by Meister et al. [64]. With the aim of analyzing the accuracy of KinectFusion, the authors scanned three scenes: a statue, a specially designed targetbox with several geometric objects, and an office room. The experiments were performed with a Microsoft Kinect. The accuracy was evaluated through both the computation of the minimal distance between each vertex of the KinectFusion point cloud and the next face of the ground-truth mesh, and the comparison of the normal of each vertex of the KinectFusion point cloud with the normal of the closest vertex in the ground-truth point cloud.

The statue was scanned at $\approx$ 100 cm distance, with a voxel side length of $\approx$ 1.6 mm; its ground truth was obtained with a Breuckmann smartSCAN-HE. The target box was scanned with a voxel side length of $\approx$ 2.7 mm. Its ground-truth measurements were manually obtained. The office was scanned from $\approx$ 100 cm to 200 cm, with a voxel side length of $\approx$ 13.7 mm. Its ground-truth data was obtained with a LiDAR using a Riegl VZ-400 time-of-flight scanner. The results of the statue and the target box showed an accuracy of 10 mm. However, highly curved and concave details below 10 mm were not well represented. It also seemed that the algorithm underestimated the volume of curved regions. The office presented errors below 80 mm for most vertices, and a median error of 36 mm. Bueno et al. performed a metrological evaluation of KinectFusion in [29], where a standard artifact composed by five delrin spheres and seven aluminum cubes was used. The accuracy and precision were evaluated at different resolutions and distances. The experiments were performed with a Microsoft Kinect. The sensor was first placed centered in front of the artifact; then, it was moved horizontally in order to obtain a more complete model of the artifact. Data at 50 cm, 100 cm, 200 cm, and 300 cm with different resolutions were acquired. In order to evaluate the accuracy, primitive spheres were fitted to the acquired point cloud as well as a primitive plane to the top face of the largest cube. The accuracy was obtained through the comparison of the diameter of the fitted spheres and the ones of the artifact. The standard deviation of the least square fitting of the geometric primitives was used to evaluate the precision. The results showed that high resolutions and close distances produce high precision. The accuracy decreased with the increment of the distance, but it was not affected by the resolution. The authors found out that according to the results, the best distance to scan the artifact is at 100 cm. Additional studies regarding KinectFusion and KinFu accuracy can be found in [71, 74].

6.2 Software and Sensors

6.2.1 KinFu

Algorithms like KinectFusion presents advantages in comparison to other reconstruction methods, such as speed and easy usage. Therefore, it is an algorithm that can be used by both researchers and end users. The software runs on a GPU, and it is divided into four parts: 1. Depth map conversion, 2. Camera tracking, 3. Volumetric integration, and 4. Raycasting. For more details see Section 3.3.2.

In this study, KinFu was used instead of KinectFusion for its flexibility regarding sensors and offline performance. KinFu's reconstruction rate, as well as the resolution of the final 3D reconstruction depend on the graphic card. The general recommendation to work with this software is to have a graphic card with at least 1 GB of memory.

6.2.2 CloudCompare and MeshLab

CloudCompare [2] is an open-source project to process 3D point clouds and triangular meshes. It contains algorithms to perform registration, resampling, comparison between point clouds and between point clouds and meshes, curvature and density computation, among others. MeshLab is an open-source system mainly developed to process and edit 3D triangular meshes. However, 3D point clouds are also accepted. This software allows filtering, texturing, and 3D mesh generation (from point clouds), among other processes.

During the experiment presented in this chapter, CloudCompare was used for the accuracy per point, curvature accuracy, and 3D reconstruction number of points evaluations (Sections 6.4.4, 6.4.5, and 6.4.6). Meanwhile, Meshlab was employed for the success rate and qualitative evaluations, as well as for manual filtering during the preparation of the data for evaluation (Sections 6.4.1, 6.4.2, and 6.4.3).

6.2.3 Sensors

Six RGBD sensors that are widely used in fields like robotics and computer vision were used (see Fig. 6.2). Such sensors acquire depth data using different technologies, such as structured light, time of flight, and stereoscopic with laser enhancement.

The sensors with structured light technology used in the experiments are: 1. Asus Xtion PRO LIVE (XT), 2. Orbbec Astra S (AS), 3. Microsoft Kinect Xbox 360 (KX), and 4. Microsoft Kinect in near mode (from now on called Kinect near mode or KN). The Microsoft Kinect v2 (K2) works with time-of-flight technology. Finally, the Intel RealSense R200 (R2) uses a stereo vision system and a laser projector to capture the depth information. More details about the operation of the aforementioned sensors can be found in Chapter 2.

6.3 Generation of the 3D Reconstructions

The 3D reconstructions acquired with KinFu and the six different sensors were analyzed as a function of the distance from the sensor to the object, the reconstruction rate, and the amount of input point clouds. Therefore, the process to obtain the reconstructions was as follows:

1. In a room, with window rollers and dark blue curtains, illuminated by three yellow incandescent ceiling lamps, each of the four objects, a dummy head (35.0 cm height), a soccer ball (11.0 cm radius), an American football (12.7×22.2×12.7 cm), and a rubber duck (23.0×21.0×28.0 cm), was placed parallel to the sensor at a determined distance using a custom-made structure (see Fig. 6.3); a support cylinder was used as a base for the American football and the soccer ball to avoid displacements. The distance was measured with respect to the dummy head's nose tip, the

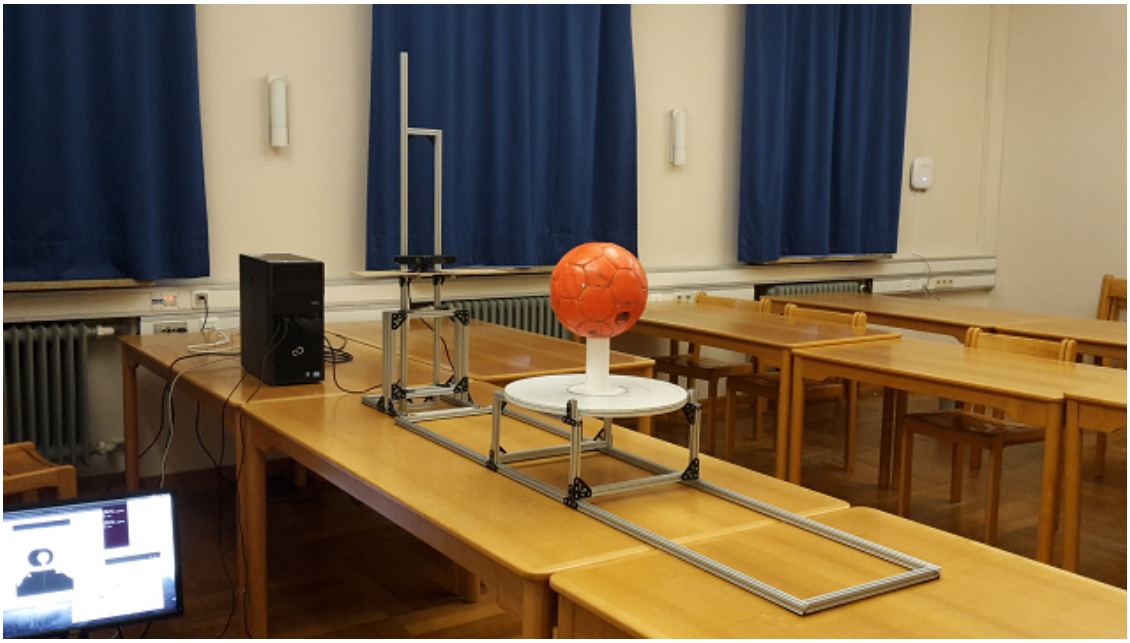


Figure 6.3: A custom-made structure with a manual rotating table was used as the setup for the acquisition of input point clouds. For the placement of the American football and the soccer ball, a support cylinder was employed.

side face of the American football, i.e., the longest part of the ball, and the beak for the rubber duck. The used distance range is shown in Table 6.1.

Sensor	Minimum [cm]	Maximum [cm]
Xtion PRO LIVE	60	120
Astra S	40	120
Kinect Xbox 360	60	120
Kinect near mode	50	120
Kinect v2	60	120
R200 (American football)	60	80
R200 (other objects)	60	120

Table 6.1: Distance range used in this experiment.

The ranges were established considering the quality of the point clouds of each object up to 120 cm (the minimum value was determined only with the dummy head). For ranges starting at 40 cm and 60 cm, the distance increased by steps of 20 cm. The Kinect near mode had an initial increment of 10 cm followed by steps of 20 cm. The shortest distance of each sensor was tested to rectify the manufacturers' data. Moreover, the largest available distance (120 cm) was set after a visual

pre-evaluation of the objects' details loss.

2. Using a manual turning table, each object was rotated while the sensor remained still. Four sets were recorded per distance for each object; each set contains the point clouds recorded in ≈ 1 revolution. The point clouds were acquired using ROS Indigo [13] packages: the sensors Xtion Pro Live, Kinect Xbox 360, and Kinect near mode used openni_launch (version: 1.9.8) [66, 65], astra_launch (version: 0.1.0) [96] was used by the Astra S, the Kinect V2 used IAI Kinect2 (version: 0.0.1) [102], and the R200 worked with RealSense (version: 2.0.3) [9].

3. With the aim of reconstructing only the object, the point clouds were automatically filtered, removing everything (e.g. tables and chairs) but the object and the back wall. In the case of the American football and the soccer ball, given their curvature at the bottom, part of the support cylinder remained due to the object-cylinder overlap.

4. The reconstructions were obtained offline for each set of point clouds using the KinFu application available in PCL. This application was solely modified to work with uncolored point clouds and to automatically save the 3D reconstructions as point clouds and meshes. The graphic card used was a Nvidia GeForce GTX 970 with 4 GB of memory. The resolution was set to 512 voxels per axis. In order to resemble situations than researchers and end users might face, such as: 1. low and full reconstruction rate, 2. various distances between the sensor and the object to reconstruct, 3. uneven amount of input point clouds, and 4. low number of input point clouds, four types of reconstructions were performed:

    - Case 1: reconstruction at 30 fps of the four objects using all the point clouds obtained during one revolution. The amount of point clouds per set ranged between 846 and 2636.
    - Case 2: reconstruction at 8 fps of the dummy head and the American football using all the input point clouds obtained during one revolution. The amount of point clouds per set ranged from 846 to 2501.
    - Case 3: reconstruction at 30 fps of the dummy head, and the American football using 800 input point clouds, distributed along one revolution.
    - Case 4: reconstruction at 8 fps of the dummy head, and the American football using 800 input point clouds, distributed along one revolution.

    For cases 2 to 4, the dummy head and the American football were selected as a representation of objects with many and few features, respectively. Furthermore, considering the minimum number of point clouds acquired in one revolution, cases three and four use 800 clouds representing few input point clouds.

6.4 Evaluations

In order to evaluate the influence of the six different RGBD sensors on the 3D reconstructions generated by KinFu, we analyzed five aspects: 1. success rate, 2. quality, 3. accuracy, 4. curvature, and 5. number of points of the reconstructions. Additionally, with the aim of investigating the impact of the reconstruction parameters (reconstruction rate and amount of input point clouds), comparisons between the reconstruction cases were also performed. Such comparisons are:

- Case 1 compared to case 2, to investigate the influence of the reconstruction rate when the point clouds recorded in $\approx$ one revolution are used.
- Case 1 compared to case 3, to examine the influence of the amount of input point clouds when the reconstruction rate is 30 fps.
- Case 3 compared to case 4, to investigate the effect of the reconstruction rate when the reconstruction is performed with a low number of input point clouds (800).
- Case 2 compared to case 4, to examine the effect of the amount of input points cloud when the reconstruction rate is 8 fps.

Since the data was collected over a time frame of $\approx$ 9 months, the accuracy and curvature were analyzed only on the dummy head and the rubber duck due to the variability of the air in the balls. Therefore, for these evaluations, the number of point clouds per set ranged from 846 to 2636 for case 1, and from 846 to 1576 for case 2. In this experiment, an initial relation between the number of point clouds per set in cases 1 and 2 and the evaluations here presented was not found. Thus, the efforts are focused on the other aspects, delegating a more in-depth analysis of this correlation for future work.

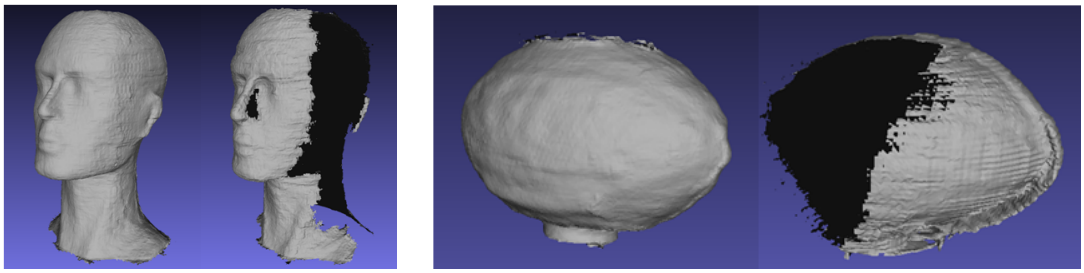
Large tables and figures of this section are shown at the end of the chapter to avoid extensive interruptions in the text.

6.4.1 Success rate

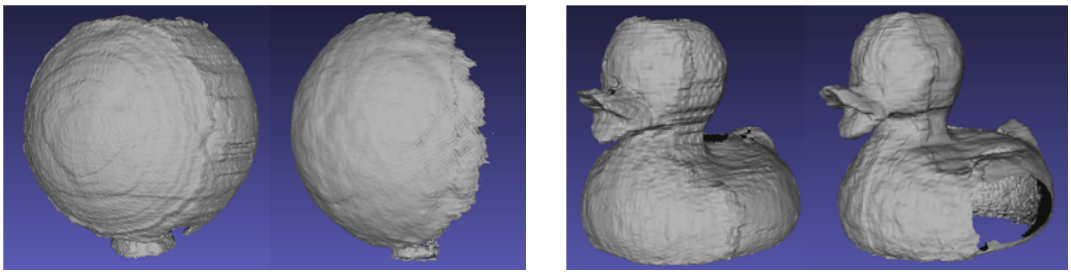
The success rate was calculated with the four reconstructions taken at each distance for each object. Therefore, for each sensor-distance, it was computed following the equation

$$SR = GR/TR\,, \tag{6.1}$$

where SR is the success rate of a specific sensor at a particular distance, GR is the amount of good reconstructions given by the sensor at the determined distance, and TR is the total number of reconstructions per sensor-distance, i. e. four. The classification as a good reconstruction was performed visually using the point cloud and the mesh obtained for each object at each distance. A reconstruction was considered successful if it presented a closed loop as shown in Fig. 6.4.



(a) Dummy head. (b) American football.



(c) Soccer ball. (d) Rubber duck.

Figure 6.4: Good and bad reconstructions. For each object, an example of a good reconstruction is shown at the left followed by a bad reconstruction exemplar at the right.

Table 6.5 and Table 6.6 show the success rate obtained in the four reconstruction cases. Objects with few or uniform features, such as the soccer ball and the American football, were not suitable for a 3D reconstruction for most of the sensors at any distance. The sensor Xtion Pro Live showed a success rate larger or equal to 0.5 at 60 cm in the four reconstruction cases for all the objects, except for the soccer ball. The Astra S presented an outstanding behavior when reconstructing the soccer ball at 40 cm and 60 cm, with a success rate of 1.0. Among all the distances and sensors, the RealSense R200 was the only one with a success rate of 1.0 for all the reconstructed objects, such rate was obtained at 60 cm in the four reconstruction cases. The sensors Kinect near mode and Kinect Xbox 360 exhibited a success rate of zero for the soccer ball at all distances. However, they reconstructed at least one time all the other objects until 80 cm. The success rate of these sensors for the American football at 100 cm and 120 cm decayed to zero. The Kinect v2 showed the lowest performance since it only reconstructed the dummy head and the duck at its closest distance. Moreover, the dummy head was the only object reconstructed beyond 60 cm by this sensor; even so, it was not reconstructed at all at 120 cm, the largest evaluated distance.

Besides the characteristics of the object to scan, the distance between the sensor and such object influenced the success reconstruction rate: the longer the distance, the lower the success rate. Moreover, after comparing the four reconstruction cases, a significant leverage of the reconstruction rate and the amount of input point clouds on the success

rate was not found. Among the four comparisons performed, only the Astra S exhibited a change of more than 25 % when reconstructing the dummy head. Such change was found when comparing the case 2 to the case 4, a decrement on the success rate at 120 cm was observed when the amount of input point clouds decreased. For the American football, the differences above 25 % were shown by both the Xtion Pro Live at 60 cm and the Kinect Xbox 360 at 80 cm when comparing cases 3 and 4 (both sensors presented a success rate increment when the reconstruction rate was reduced to 8 fps), and by the Astra S at 80 cm when comparing cases 2 and 4 (the success rate decreased when the amount of input point clouds decreased).

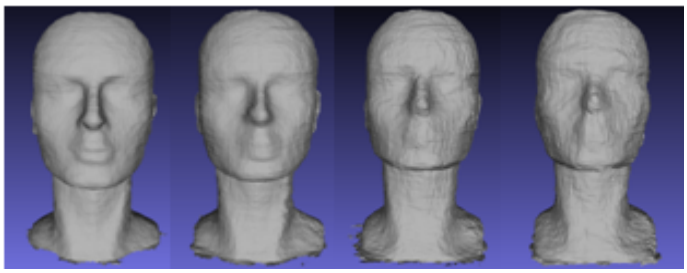
6.4.2 Qualitative evaluation

The visual quality of the reconstruction decreased with the distance, as is shown in Fig. 6.5 and Fig. 6.6, due to the reduction of points of each reconstruction (see Section 6.4.6). For all sensors, the most similar reconstruction to the object was obtained at the shortest distance. When the distance increased, artifacts on the smoothest parts of the objects (mainly on the dummy head) were observed for all the sensors but the R200 and the Kinect v2, as is shown in Fig. 6.7a. Moreover, artifacts on the top of the reconstructions were expected since the top of the objects was not scanned. However, the reconstructions with the R200 presented a crown-shaped artifact, particularly visible on the dummy head (see Fig. 6.6c); this feature was not shown by any other sensor. Additionally, the Kinect v2 exhibited a pointy reconstruction of the rubber duck's eyes, as shown in Fig. 6.7b. This particular behavior may be due to the black paint on the duck's eyes. Other artifacts also appeared in the reconstructions, ranging from small and medium sizes to large as shown in Fig. 6.7c, Fig. 6.7d, and Fig. 6.7e. Furthermore, some reconstructions exhibited a complete or partial double point cloud effect, i.e., there is an additional complete or partial second point cloud inside the external one (see Fig. 6.7f). The large and double point cloud artifacts are quantified in Table 6.2. From this quantification, it is important to highlight that: 1. the number of amiss reconstructions given by the Kinect near mode is significantly higher in comparison to the other sensors in all the cases, and 2. the Kinect v2 and the R200 did not present the evaluated artifacts in any of their reconstructions.

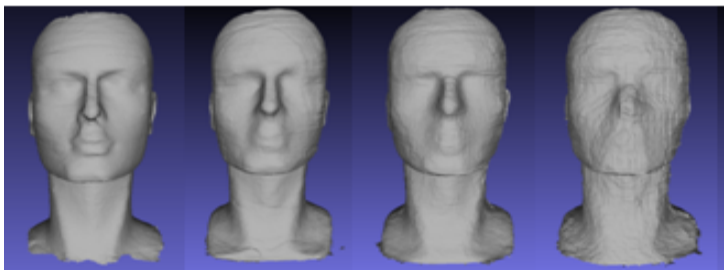
Among the reconstruction cases a significant visual difference in the reconstructions was not found. The comparison of the dummy head's reconstructions generated at the closest and furthest comparable distances with the four reconstruction cases is shown in Fig. 6.16 and Fig. 6.17.

6.4.3 Data preparation for accuracy, curvature, and number of points evaluations

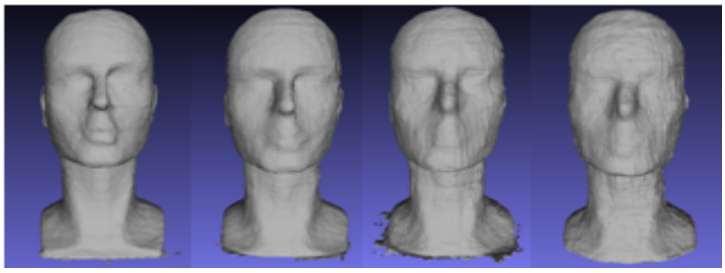
The accuracy and the curvature were evaluated through the comparison of the obtained reconstructions as point clouds to a ground-truth mesh. The mesh was acquired with an



(a) Reconstructed from 60 cm to 120 cm with point clouds obtained with the Xtion Pro Live.



(b) Reconstructed from 40 cm to 100 cm with point clouds obtained with the Astra S.

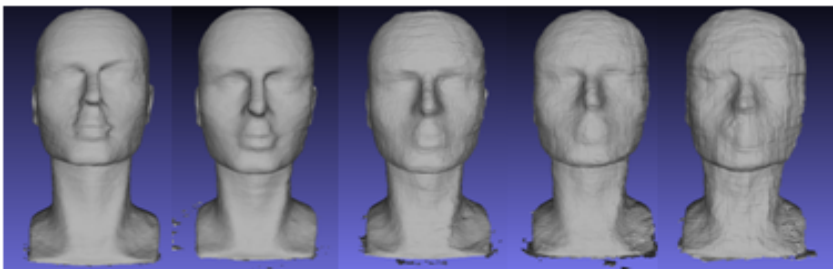


(c) Reconstructed from 60 cm to 120 cm with point clouds obtained with the Kinect Xbox 360.

Figure 6.5: Visual quality evaluation of the dummy head's reconstructions obtained with the sensors Xtion Pro Live, Astra S, and Kinect Xbox 360.

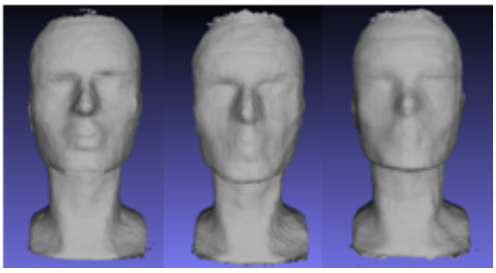
	Cases 1 & 2		Cases 3 & 4	
Sensor	30 fps	8 fps	30 fps	8 fps
Xtion	4	2	1	1
Astra S	3	3	1	0
K. Xbox 360	3	1	0	2
K. near mode	13	7	8	5
K. v2	0	0	0	0
R200	0	0	0	0

Table 6.2: Quantification of the large and double point cloud artifacts.



(a) Reconstructed from 50 cm to 120 cm with point clouds obtained with the Kinect near mode.

(b) Reconstructed from 60 cm to 100 cm with point clouds obtained with the Kinect v2.



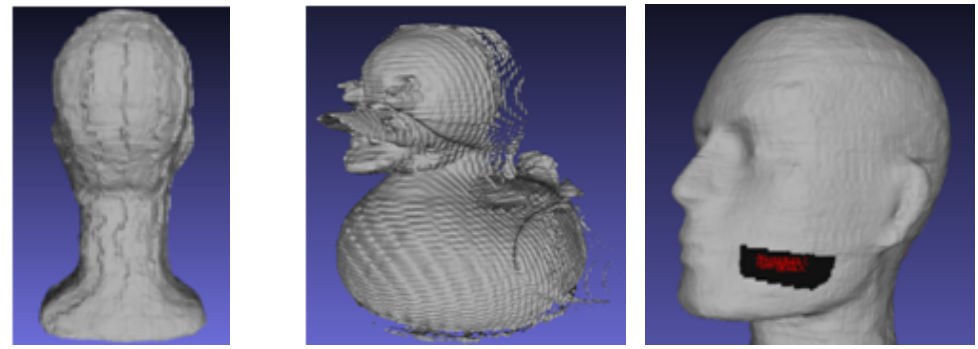
(c) Reconstructed from 60 to 100 cm with point clouds obtained with the RealSense R200.

Figure 6.6: Visual quality evaluation of the dummy head's reconstructions obtained with the sensors Kinect near mode, Kinect v2, and RealSense R200.

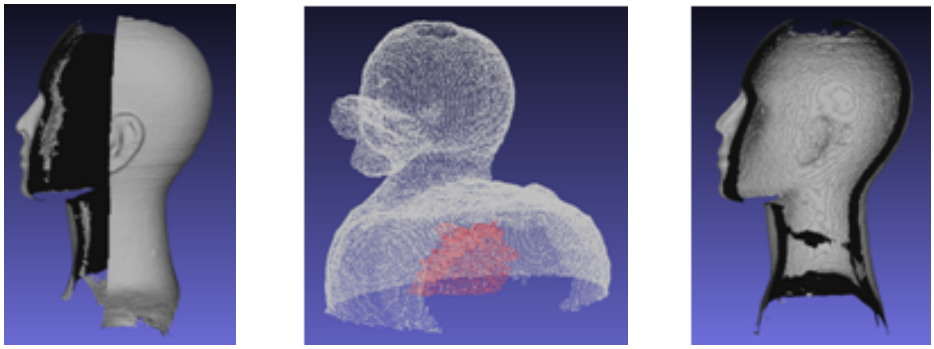
Artec EVA Scanner at the Fine-mechanical workshop in the Max Plank Institute for Biological Cybernetics. This sensor has a 3D resolution up to 0.5 mm, a 3D point accuracy up to 0.1 mm, and a 3D accuracy over distance up to 0.03 % over 100 cm. Figure 6.8 and Table 6.3 show the ground-truth meshes and their characteristics.

Model	Vertices	Faces
Dummy head	304.767	609.522
Soccer ball	110.573	221.142
American football	115.757	231.510
Rubber duck	277.686	555.368

Table 6.3: Number of vertices and faces of the ground-truth models.



(a) (b) (c)



(d) (e) (f)

Figure 6.7: Artifacts in the reconstructions.(a) Artifact on the back of the head produced by the sensors with projection-pattern technology. (b) The Kinect v2 produced a pointy effect of the duck's black eyes. (c) In red, a small artifact. (d) Vertical large internal artifact. (e) In red, a large artifact. (f) Vertical cut view of a dummy head with double point cloud effect. Figure best viewed in color.

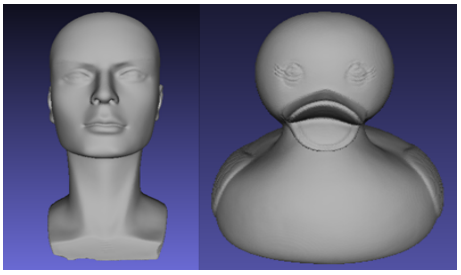


Figure 6.8: Ground-truth 3D models obtained with the Artec EVA Scanner. Left: dummy head, right: large rubber duck.

To analyze the accuracy, each reconstruction as point cloud was registered to the ground-truth mesh. This process was performed using the CloudCompare's iterative closest point tool on all the points of each point cloud. The RMS difference was set to $1e^{-8}$, close to the CloudCompare's computation accuracy limit. Point clouds' outliers

were manually removed using MeshLab to avoid their influence on the registration.

For the accuracy and curvature evaluations, only the collections of four reconstructions with success rate larger or equal to 0.75 were used. For those whose success rate was equal to 0.75, the worst value was duplicated; thus, all the compared collections consist of four reconstructions.

Since the perspective and distance of the sensors with respect to the objects as well as the automatic segmentation affect the volume of each object to be reconstructed, the ground-truth meshes and the 3D point clouds were cropped at the bottom as shown in Fig. 6.9. The cropped version of the clouds and meshes were used for the remaining evaluations.

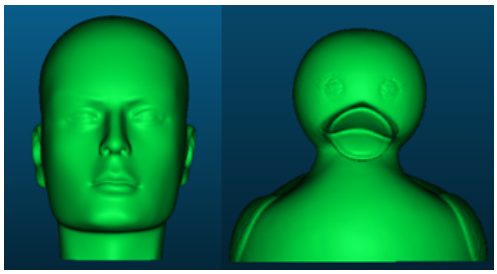


Figure 6.9: Cropped ground-truth meshes. Left: dummy head, right: rubber duck.

6.4.4 Accuracy per point evaluation

The accuracy per point was defined as the distance between each point of the reconstructed point cloud and the closest surface of the ground-truth mesh, following the equation

$$e_p = s - g \,, \tag{6.2}$$

where e_p is the difference between the reconstructions' values (s) and the ground truth's (g); it was obtained using CloudCompare. In order to summarize such distances, the median and mean per point cloud were calculated. Hence, Fig. 6.18, Fig. 6.19, and Appendix A present the median and mean absolute error per sensor-distance computed as the median and mean of the four reconstruction's median and mean values obtained per sensor-distance. The aforementioned figures are presented at the end of this chapter to avoid large interruptions on the text.

The absolute median error range of the sensors varied between 0.8 mm and 4.9 mm, meanwhile the absolute mean error ranged from 0.9 mm to 5.3 mm. The results are acceptable accuracy values for most applications. The sensors' rank (from most to less accurate) presented a similar behavior for the four reconstruction cases. When the object was placed at 60 cm from the sensors, the Astra S exhibited the best behavior with a median error of 0.9 mm, followed by the Xtion Pro Live in cases 1 and 2, and for the

RealSense R200 in cases 3 and 4 (when reconstructing the rubber duck, the Astra S, the Kinect near mode, and the R200 were tied in the first place). At 80 cm and 100 cm, the first two places were mostly alternated between the RealSense R200 and the Kinect near mode with errors between 0.8 mm and 1.3 mm. The sensors Kinect near mode and Kinect Xbox 360 presented the lowest errors when the object was located at 120 cm away from the sensor. The largest errors were given by the Kinect v2 at all distances and in all cases in which the Kinect v2 was evaluated; the reconstruction of the rubber duck at 60 cm presented the highest median and mean errors: 4.9 mm and 5.3 mm, respectively.

After performing the comparison between reconstruction cases (case 1 vs case 2, case 1 vs case 3, case 3 vs case 4, and case 2 vs case 4), the results showed that the reconstruction rate and the amount of input point clouds did not considerably influence the error of each sensor. For instance, accuracy changes equal or higher than 0.5 mm were produced only by the sensors Astra S, Xtion Pro Live, and Kinect Xbox 360. At 120 cm, the Astra S only reconstructed the dummy head at 8 fps and using all the point clouds recorded in one revolution (case 3). Additionally, while comparing cases 1 and 3, the Xtion Pro Live (at 60 cm and 120 cm) and the Kinect Xbox 360 (at 120 cm) presented an error increment when the reconstruction rate decreased. Furthermore, when comparing the cases 2 and 4, the sensor Xtion Pro Live also presented a higher error when the reconstruction rate decreased.

The signed median errors (see Fig. 6.20, Fig. 6.21, and Appendix A) determined that the sensors Xtion Pro Live, Kinect Xbox 360, and Kinect v2 tended to underestimate the objects' dimensions. In contrast, the RealSense R200 had a tendency to overestimate them. The sensors Astra S and Kinect near mode presented a mixture of under- and overestimation. The Astra S underestimated the size of the objects when these were located at 40 cm from the sensor. From 60 cm onward the objects' measurements were overestimated. The Kinect near mode's trend was to underestimate the dummy head's dimensions when it was located at up to 80 cm; for the rubber duck, the underestimation was only at 50 cm. At larger distances, the Kinect near mode was prone to overestimate the measurements.

Comparison of the presented results to the state of the art

In order to compare the presented results with the literature, the focus is on the reconstruction of the dummy head in case 1 for the Kinect Xbox 360 and the Kinect near mode. Both sensors exhibited better accuracy than the one shown by Meister et al. [64] and Bueno et al. [29], and it is similar to the one reported by Pagliari et al. in [74] for the Lego bricks (see Table 6.4). However, a direct comparison with [64], Bueno et al. [29], and [74] (when reconstructing the statue) is not recommended since: 1. in [29] sensor calibration details were not given, 2. in [64] the camera calibration was not performed, and 3. in [74] the data of the statue was acquired with the sensor in an upper position looking downwards from multiple view angles without considering the implications of such angles.

Approach	Abs. median [mm]	Abs. mean [mm]	Signed median [mm]
Mesiter's	4.5 mm (statue)	6.6 mm (statue)	-
	5.6 mm (box)	9.3 mm (box)	-
Bueno's	$\approx$ 2.0 to 4.0 mm	-	-
	from 50 to 120 cm		
Pagliari's	-	-	$\approx$ 1.8 mm (statue)
			≈ -1.8 mm (Lego bricks)
Presented Experiment			
Kinect Xbox 360	1.1 to 2.3 mm	1.3 to 2.6 mm	-2.6 to 0.2 mm
Kinect near mode	0.9 to 2.1 mm	1.1 to 2.8 mm	-2.6 to -0.9 mm

Table 6.4: The accuracy per point of the dummy head in case 1 is compared to the accuracy reported in the state of the art.

6.4.5 Curvature accuracy evaluation

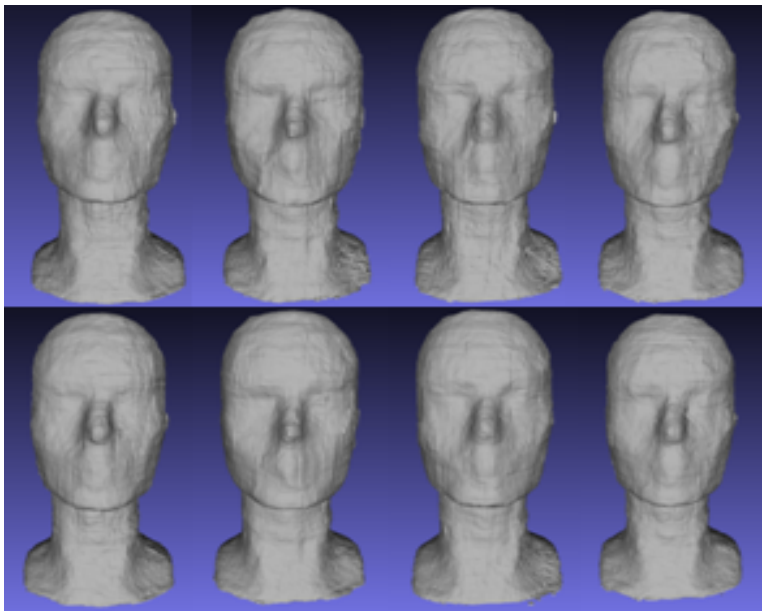
The curvature is evaluated to quantify the reconstructions' detail loss when the distance to the sensor increases. Therefore, the difference between the median and mean of the mean curvatures of each reconstruction's point cloud and the median and mean of the mean curvatures of the ground-truth mesh were calculated following the equation

$$e_v = g - s \,, \tag{6.3}$$

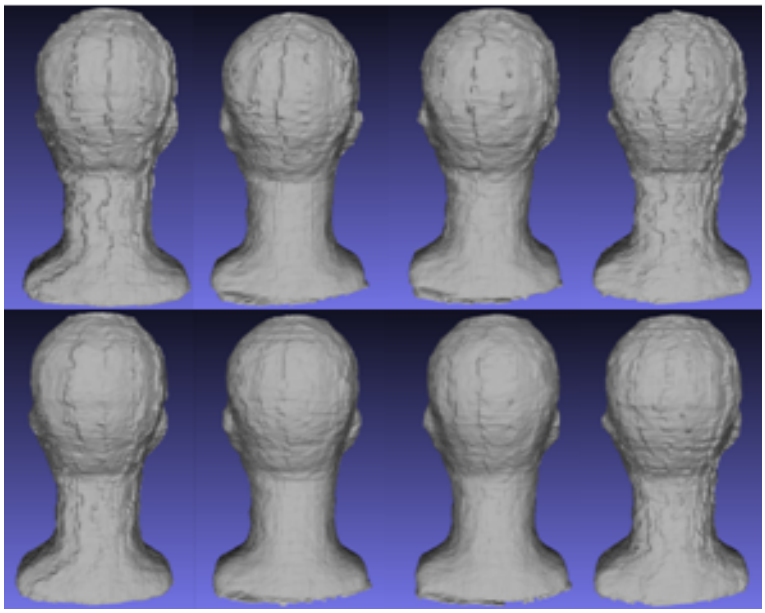
where e_v is the difference between the ground truth's (g) and the reconstructions' (s) values. The curvatures were computed in CloudCompare; since CloudCompare calculates curvatures only on point clouds, the curvatures of the ground truths were calculated using the vertices of the meshes. The curvatures were obtained using a kernel of 1.5 cm, which was empirically obtained looking at curvatures at 120 cm. Fig. 6.22, Fig. 6.23, and Appendix B present the relative unsigned median and mean error per sensor-distance computed as the unsigned median and mean of the four reconstruction's difference values obtained per sensor-distance.

The median detail loss was between 0.9 % and 95.7 %; meanwhile the mean curvature error ranged from 0.8 % to 52.7 %. For most sensors the mean curvature error followed a quadratic function that increased with the distance between the sensor and the object, this increment was produced by the reduction of the number of points of each reconstruction (see Section 6.4.6). Among all reconstruction cases, the sensors Kinect near mode and Kinect v2 presented the first- and second-best results when reconstructing the dummy head at 60 cm; regarding the rubber duck, the Kinect Xbox 360 occupied the first place followed by the Kinect v2. At 80 cm and 100 cm, the Kinect v2 was in the first or second position in all cases; this sensor showed an outstanding behavior at 100 cm. However, since the rubber duck was not reconstructed by the Kinect v2 at these distances, the first place was occupied by either the Xtion Pro Live or the Kinect Xbox 360 when this object was evaluated. At 120 cm, the two first places were held by the Xtion Pro Live and the

Kinect Xbox 360. The rubber duck was not reconstructed by the Xtion Pro Live; thus, the first and second place were occupied by the Kinect near mode and the Kinect Xbox 360. The worst results were exhibited by the Astra S and the RealSense R200.



(a) Frontal view.



(b) Back view.

Figure 6.10: Dummy head's reconstruction at 120 cm in case 1 and 3.

After performing the comparison between the reconstruction cases, the results showed that the reconstruction rate and the amount of input point clouds did not significantly influence the visual detail loss of each sensor. For instance, Fig. 6.10 depicts the very similar reconstructions given by the sensor Kinect Xbox 360 in cases 1 and 3 (only a small increment on the smoothness is perceived in case 3). The curvature median error increased 13.2 % when few input point clouds were used (case 3); such value is the largest difference between cases.

The signed curvature median errors (see Fig. 6.24, Fig. 6.25, and Appendix B) showed that all the sensors tend to underestimate the curvatures. However, the Kinect near mode overestimated the curvatures of the dummy head when all the point clouds recorded in one revolution at 60 cm were used. Additionally, the curvatures of the aforementioned object were overestimated also by the Kinect Xbox 360 at 60 cm during the reconstruction case 2. The rubber duck presented overestimated curvatures generated on the reconstructions given by the Astra S and the Kinect v2 at 40 cm and 60 cm, respectively.

6.4.6 Number of points

The median of the number of points of the dummy head's point clouds at each distance in the four reconstruction cases is shown in Fig. 6.11. We observed that: 1. the number of points decreased with the distance following a quadratic function, as expected, 2. the number of points per distance was independent of the sensor technology, as well as of the reconstruction case. For reference, the used depth image's resolution of the Xtion Pro Live, the Astra S, the Kinect near mode, and the Kinect Xbox 360 was 640×480; the Kinect v2 had a resolution of 512×424. Moreover, the resolution of the RealSense R200 was 480×360

6.4.7 Sensor Intel Realsense D415

At the time of writing this manuscript, a new state-of-the-art sensor was launched by Intel: the Intel RealSense D415, which works with the same technology as the RealSense R200. Therefore, a preliminary study of the dummy head's reconstructions generated with point clouds taken with a D415 was performed. Evaluations of the success rate, the quality, the accuracy per point, and the curvature accuracy were executed. The data acquisition was performed from 30 cm to 120 cm, the first step was of 10 cm succeeded by steps of 20 cm. The data was obtained following the process described in Section 6.3: the dummy head was placed on a manual rotating table and four sets of point clouds were acquired. Each set contains the point clouds obtained in ≈ 1 revolution. The amount of point clouds per set ranged from 1421 to 2812.

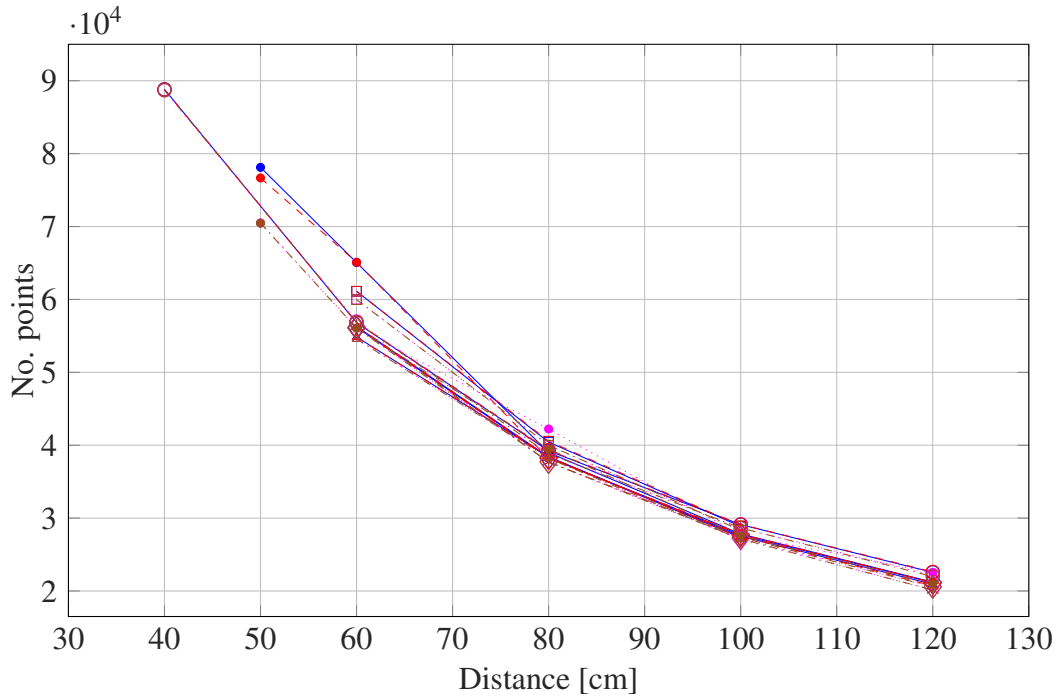


Figure 6.11: Median of the number of points of the dummy head's point clouds. The sensors are represented by: ∗ Xtion Pro Live, ○ Astra S, ● Kinect near mode, ◇ Kinect Xbox 360, □ Kinect v2, and △ RealSense R200. Solid blue, dashed red, dotted magenta, and dash-dotted brown lines represent the reconstruction cases 1 to 4, respectively.

Success rate

The success rate was calculated as specified in Section 6.4.1 and is depicted in Fig. 6.12. The success rate among the reconstruction cases was the same only at 40 cm and 60 cm, 0.5 and 1.0, respectively. The reconstruction cases 1 and 2 (reconstructions with all the point cloud acquired in 1 revolution) exhibited a larger success rate at 30 cm than cases 3 and 4 (reconstruction with 800 input point clouds). Conversely, at 80 cm, 100 cm, and 120 cm the cases 3 and 4 presented larger success rates than cases 1 and 2. Success rates equal or larger than 0.75 were obtained at 30 cm only by cases 1 and 2, and at 80 cm and 100 cm by cases 3 and 4; at 40 cm and 120 cm none of the reconstruction cases reached such value.

Qualitative evaluation

The visual quality evaluation of the reconstructed dummy head depicted horizontal artifacts on the surface, as shown in Fig. 6.13. Such artifacts might be generated by the interaction of the infrared rays with the left side of the dummy head's surface as shown in Fig. 6.14, where horizontal lines (whose length decreases when the distance from the

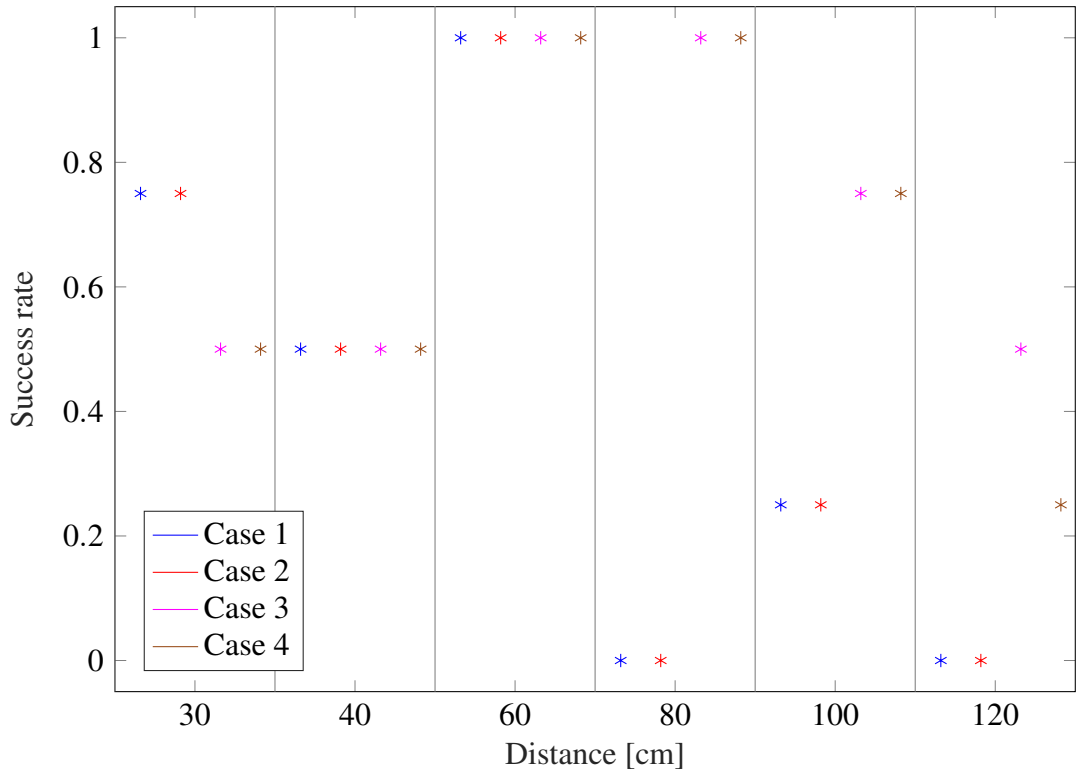
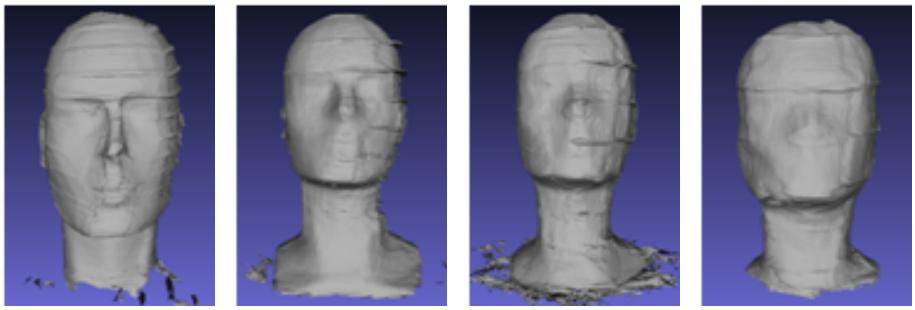


Figure 6.12: Success rate of the dummy head's reconstructions obtained with point clouds acquired with the sensor RealSense D415 in all reconstruction cases.

object to the sensor increases) are observed outside the object's surface. Internal small and large artifacts, as well, as double point clouds (as described in Section 6.4.2) are also exhibited by the reconstructions obtained with the RealSense D415.



(a) 30 cm (b) 60 cm (c) 80 cm (d) 100 cm

Figure 6.13: Artifacts on the reconstructions generated with data taken by the sensor RealSense D415.

Figure 6.14: Horizontal lines are observed at the left side of the dummy head's surface, such lines produce artifacts on the reconstructions and might be the result of the interaction of the infrared projector with the surface. From left to right: the distance between the sensor and the dummy head are 30 cm, 60 cm, and 120 cm. Frontal and side views of the object are depicted to exhibit the persistence of the lines at the left side of the dummy head.

A comparison among the reconstruction cases is only possible at 60 cm since it is the only distance at which the four reconstruction cases were evaluated. Figure 6.15 depicts the visual changes among the cases. Large differences among the reconstructions are not visible besides the horizontal artifacts, which spread along the entire face on the reconstructions obtained at cases 3 and 4.

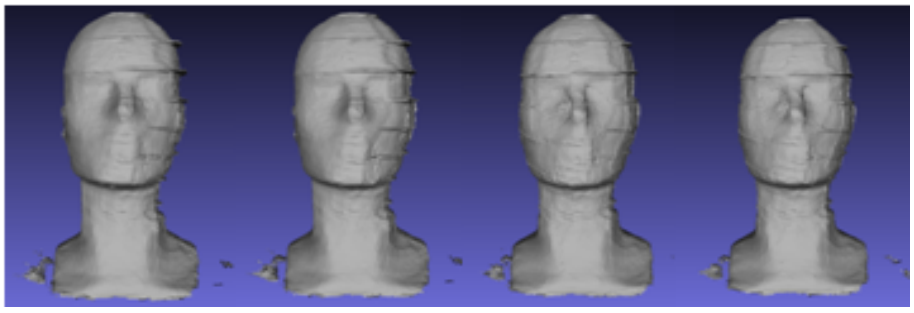


Figure 6.15: Dummy head reconstructed at 60 cm in all the reconstruction cases by the sensor RealSense D415.

Accuracy per point and curvature evaluations

The median and mean accuracy per point error was calculated as indicated in Section 6.4.4. Figure 6.26a presents the absolute median and mean values. The median error range varied between 8.2 mm and 10.1 mm. The mean error ranged from 8.1 mm to 11.4 mm. The median values are, at least, 3.3 mm larger than the largest median error found in Section 6.4.4. The signed accuracy errors (depicted in Fig. 6.26b) demonstrated that the 3D reconstructions given by the D415 are underestimated at all distances and in all reconstruction cases.

Regarding the details loss, Fig. 6.27a shows the relative unsigned median and mean curvature errors in all the reconstruction cases. These results were calculated as indicated in Section 6.4.5. The median error varies between 4.2 % and 37.5 %, which describes a similar error range than the error presented in Section 6.4.5. The signed median curvature errors (see Fig. 6.27b) showed that the D415 overestimates the curvature at 30 cm probably due to the horizontal artifacts (see Section 6.4.7). The reconstructions' curvature were underestimated at all the other distances.

The comparison among the reconstruction cases is only possible at 60 cm. At such distance, both the accuracy per point and the curvature given by the cases 1 and 2 presented lower values than cases 3 and 4. Regarding the accuracy per point, the difference is up to 1.0 mm. Furthermore, the curvature difference error is up to 4.9 % (given by case 1 and case 3).

6.5 Summary and Conclusions

In this chapter, the qualitative and quantitative results of an experiment that studies the influence of six well-known RGBD sensors on 3D reconstructions obtained with KinFu were presented. The sensors Microsoft Kinect Xbox 360, Microsoft Kinect near mode, Microsoft Kinect v2, Asus Xtion Pro Live, Orbbec Astra S, and Intel RealSense R200 reconstructed a dummy head, a soccer ball, an American football, and a rubber duck. The reconstructions were acquired resembling scenarios that researchers and end users might face: 1. low and full reconstruction rate, 2. various distances between the sensor and the object to reconstruct, 3. uneven amount of input point clouds, 4. low number of input point clouds[2]. This work provides a technically-grounded guideline for sensor selection to KinFu users. Moreover, the 3D data are available for download.

Based on the results of the presented experiments, the outcome of Kinfu depends on the sensor used to record its input and the selection of a specific RGBD sensor depends on the requirements of each application. However, the object to be scanned should have enough non-uniform features and should be placed at a distance up to 100 cm to preserve details. In case that accuracy is more important than details and if an error up to 5 mm is acceptable, any of the six tested sensors is recommended at any of the examined distances

[2]The point clouds should be overlapped to work with KinFu.

and scenarios. However, the lowest accuracy errors ($\leq$ 1.0 mm) were mostly found with the Astra S, the Kinect near mode, and the RealSense R200. Regarding the mean curvature, the larger the distance from the object to the sensor, the larger the curvature loss. Moreover, with regard to the number of points of the reconstructions, all the sensors followed a similar descending quadratic function. Therefore, the number of points of the reconstructions as point clouds is independent of the sensor technology.

The Kinect near mode and the Kinect v2 exhibited low curvature errors; however, it is important to recall that the Kinect near mode presented the highest amount of large artifacts and the Kinect v2 had the largest accuracy errors. The RealSense R200 showed a high accuracy but large curvature errors; these curvature errors likely stem from the crown-shaped artifacts created by the interaction of the rays with the object. Thus, this sensor might be recommended if a complete reconstruction is performed, i.e. the top of the object is also recorded. In order to reconstruct objects at 60 cm, the Astra S and the Xtion Pro Live are both suggested, because they presented a similar behavior regarding the accuracy (difference error up to 0.6 mm between them) and curvature; the Kinect near mode was discarded due to the number of large artifacts. At 80 cm, the Xtion Pro Live surpassed the Astra S, conserving more details. The Kinect Xbox 360 exhibited average results, its lowest accuracy errors were at 100 cm. In this study, an initial relation between the number of point clouds used to reconstruct the objects in cases 1 and 2 and the performed evaluations was not found.

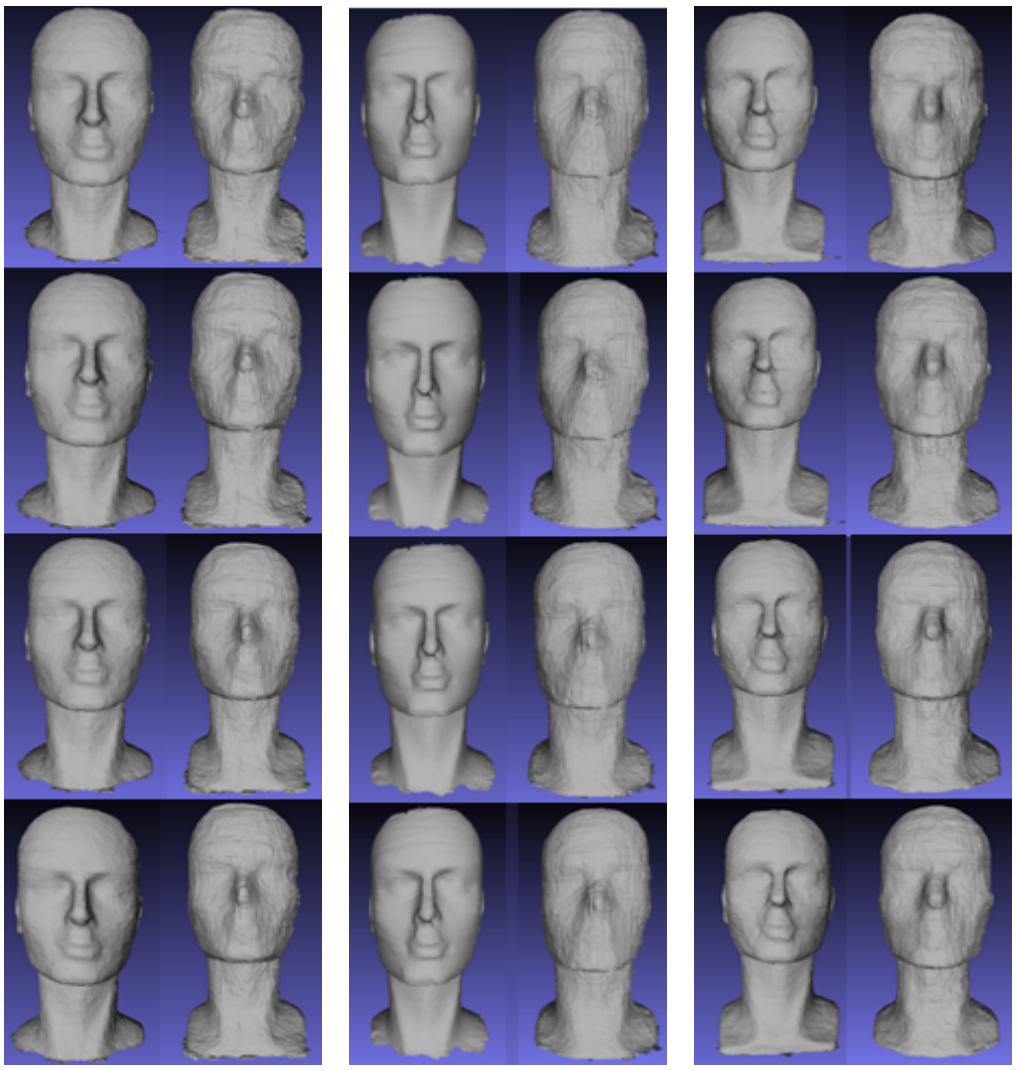
In addition to the aforementioned sensors, the state-of-the-art sensor Intel RealSense D415 was also tested through a preliminary study, in which the dummy head was the only reconstructed object. Horizontal protuberances were observed on the models. Such artifacts might be generated by the interaction of the projected laser and the dummy head's surface since they are produced by depth measurements, whose length decreased when the distance between the object and the sensor increased. Furthermore, the reconstructions were affected by the number of input point clouds used to generate them, i.e., using all the point clouds recorded in one revolution, the dummy head was mostly reconstructed at 30 cm and 60 cm. However, using only 800 input point clouds the object was mostly reconstructed at 60 cm, 80 cm, and 100 cm. The median accuracy errors given by the D415 were (at least) 3.3 mm larger than the other sensors. Nonetheless, the curvature errors presented a similar behavior than the other sensors.

	Distance (cm)	Sensor XT	AS	KX	KN	K2	R2
Case 1	40	-	**1.0**	-	-	-	-
Dummy head	50	-	-	-	**1.0**	-	-
	60 - 100	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	120	**1.0**	0.50	**1.0**	**1.0**	0	0
Soccer ball	40	-	**1.0**	-	-	-	-
	50	-	-	-	0	-	-
	60	0	**1.0**	0	0	0	**1.0**
	80 - 120	0	0	0	0	0	0
American football	40	-	0.25	-	-	-	-
	50	-	-	-	0.50	-	-
	60	0.75	0	0.50	0.75	0	**1.0**
	80	0	0.25	0.25	0.25	0	0
	100 - 120	0	0	0	0	0	-
Rubber duck	40	-	**1.0**	-	-	-	-
	50	-	-	-	**1.0**	-	-
	60	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	80	**1.0**	**1.0**	0.50	**1.0**	0	**1.0**
	100	**1.0**	**1.0**	**1.0**	**1.0**	0	0.50
	120	0.50	0.75	0.75	**1.0**	0	0
Case 2	40	-	**1.0**	-	-	-	-
Dummy head	50	-	-	-	**1.0**	-	-
	60	**1.0**	**1.0**	0.75	**1.0**	**1.0**	**1.0**
	80 - 100	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	120	**1.0**	0.75	**1.0**	**1.0**	0	0
American football	40	-	0.25	-	-	-	-
	50	-	-	-	0.50	-	-
	60	0.75	0	0.50	0.75	0	**1.0**
	80	0	0.50	0.25	0.25	0	0
	100 - 120	0	0	0	0	0	-

Table 6.5: Success rate calculated as defined in Section 6.4.1. for cases 1 and 2.

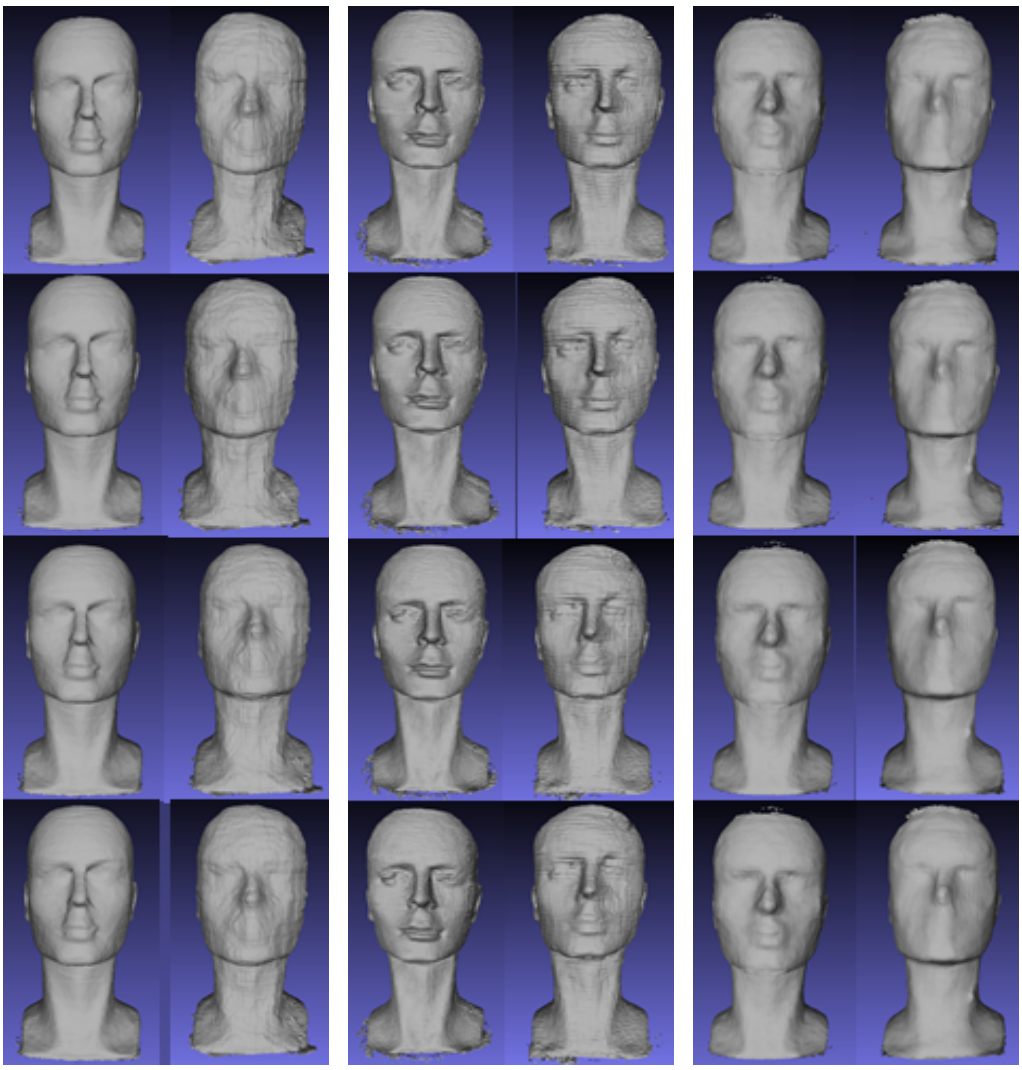
		Sensor					
	Distance (cm)	XT	AS	KX	KN	K2	R2
Case 3	40	-	**1.0**	-	-	-	-
	50	-	-	-	**1.0**	-	-
Dummy head	60	0.75	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	80 - 100	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	120	**1.0**	0.50	**1.0**	**1.0**	0	0
	40	-	0.25	-	-	-	-
	50	-	-	-	0.50	-	-
American football	60	0.50	0.25	0.50	**1.0**	0	**1.0**
	80	0.25	0.25	0	0.50	0	0
	100 - 120	0	0	0	0	0	-
Case 4	40	-	**1.0**	-	-	-	-
	50	-	-	-	**1.0**	-	-
Dummy head	60	0.75	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	80 - 100	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	120	**1.0**	0.25	**1.0**	**1.0**	0	0
	40	-	0	-	-	-	-
	50	-	-	-	0.50	-	-
American football	60	**1.0**	0.25	0.25	0.75	0	**1.0**
	80	0.25	0	0.50	0.50	0	0
	100 - 120	0	0	0	0	0	-

Table 6.6: Success rate calculated as defined in Section 6.4.1 for cases 3 and 4.



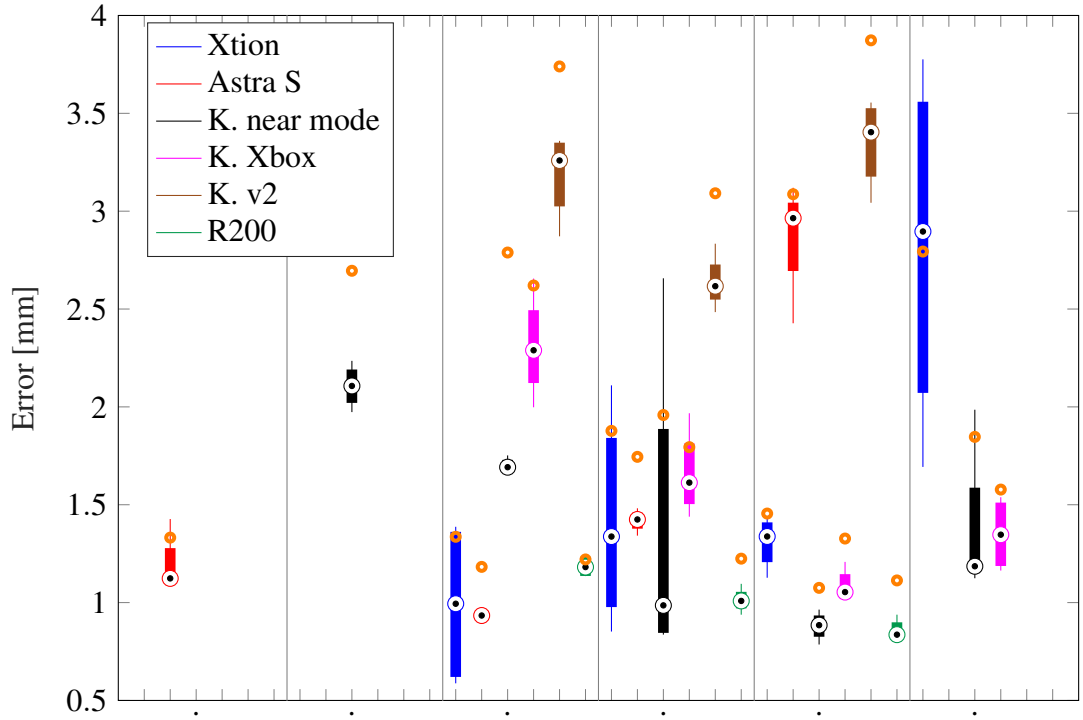
(a) (b) (c)

Figure 6.16: Visual quality evaluation among the four reconstruction cases. The rows represent cases 1 to 4. The columns are examples of the closest and furthest distance reconstructed. (a) Reconstructed with the Xtion Pro Live at 60 cm and 120 cm. (b) Reconstructed with the Astra S at 40 cm and 100 cm. (c) Reconstructed with the Kinect Xbox 360 at 60 cm and 120 cm.

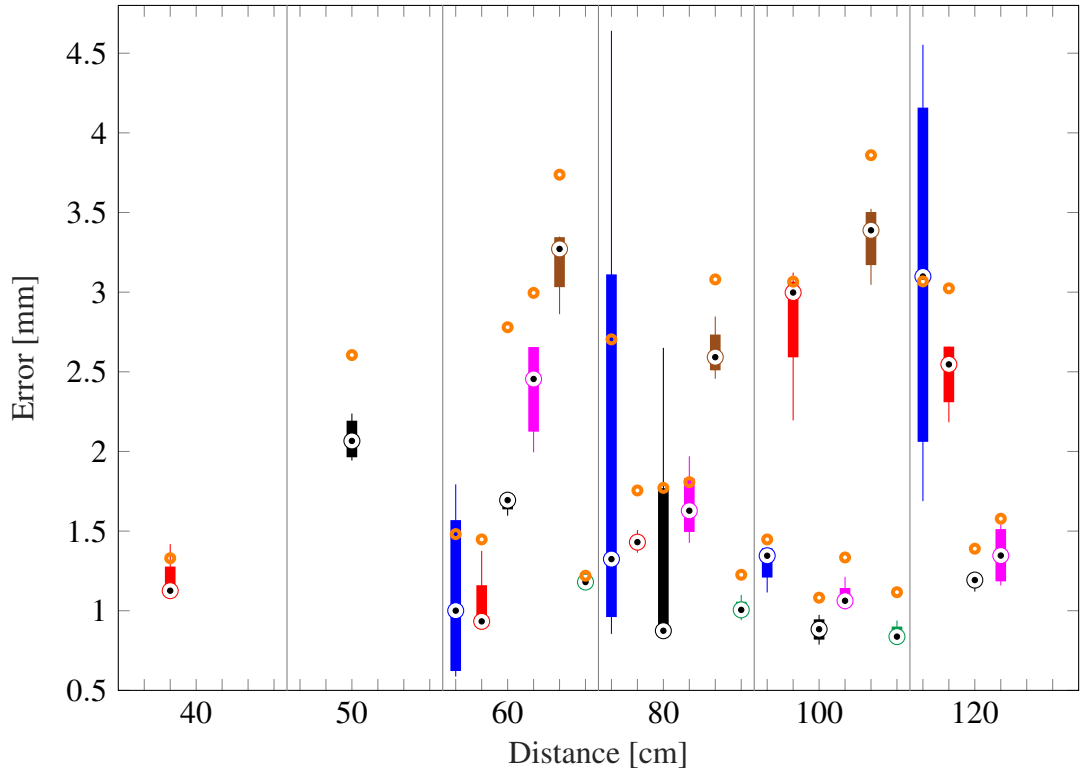


(a) (b) (c)

Figure 6.17: Visual quality evaluation among the four reconstruction cases. The rows represent cases 1 to 4. The columns are examples of the closest and furthest distance reconstructed. (a) Reconstructed with the Kinect near mode at 50 cm and 120 cm. (b) Reconstructed with the Kinect v2 at 60 cm and 100 cm. (c) Reconstructed with the RealSense R200 at 60 cm and 100 cm.

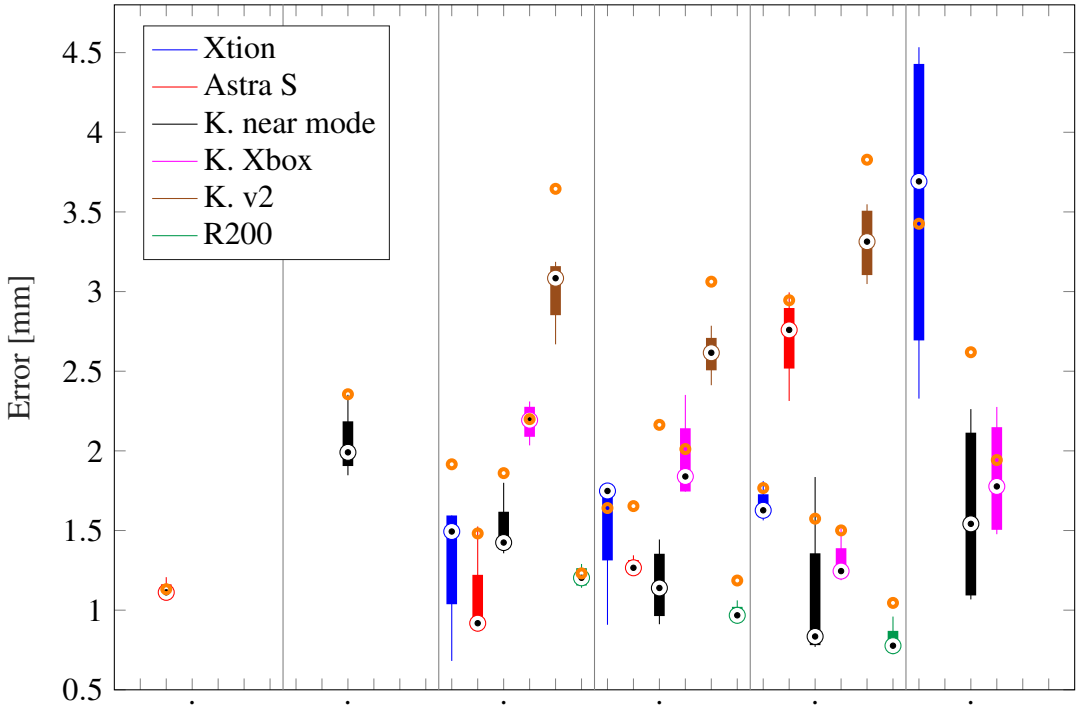


(a) Absolute error per point of the dummy head in case 1.

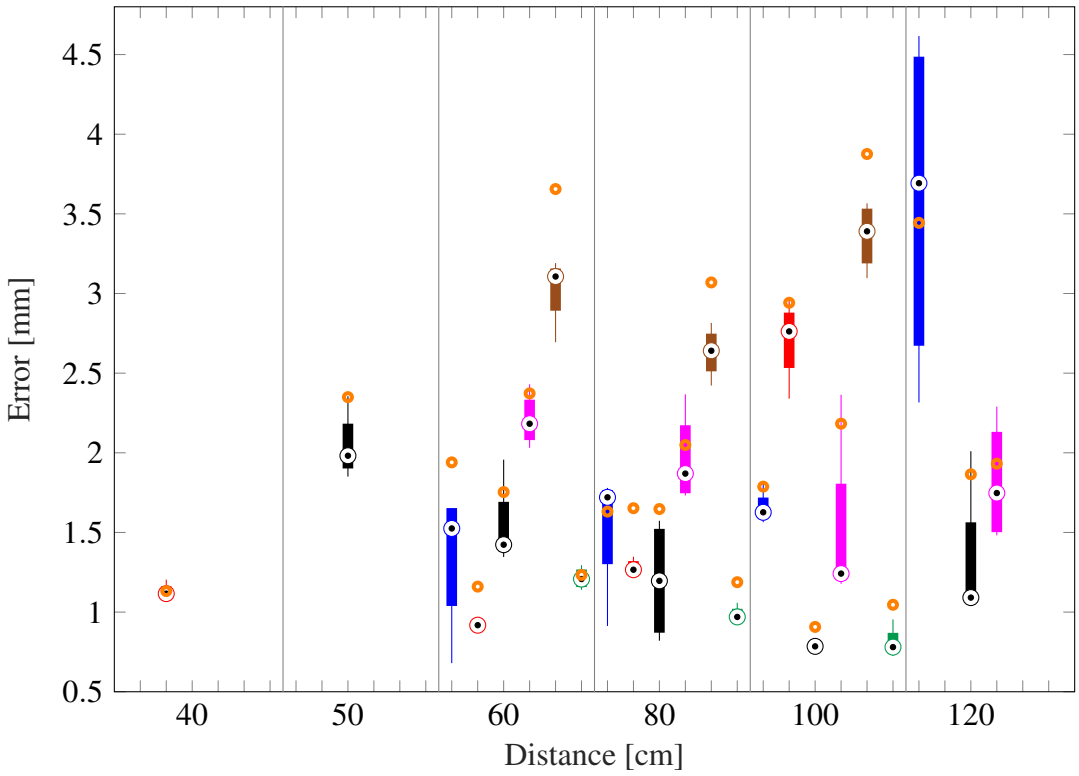


(b) Absolute error per point of the dummy head in case 2.

Figure 6.18: Absolute error per point in cases 1 and 2. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

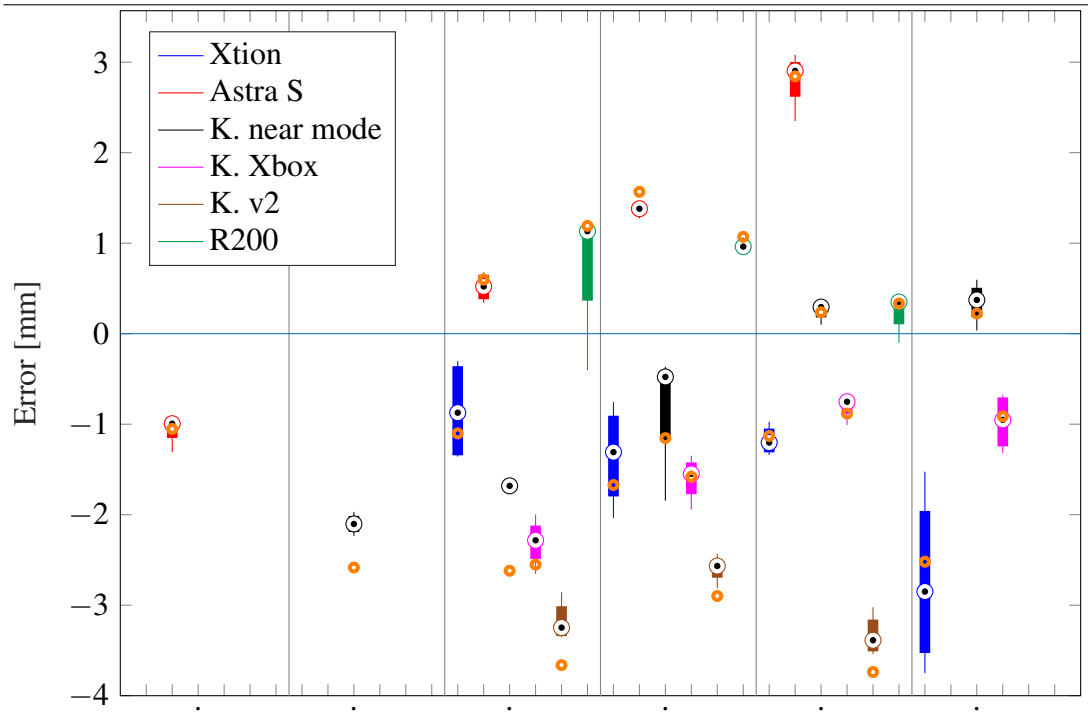


(a) Absolute error per point of the dummy head in case 3.

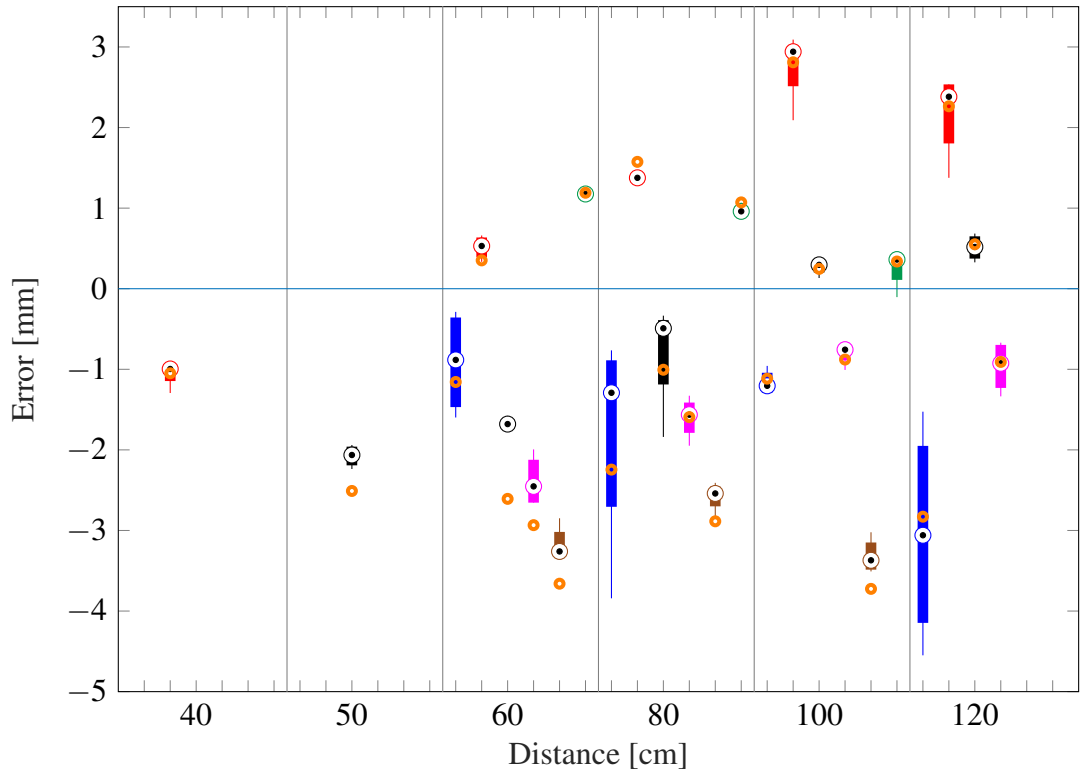


(b) Absolute error per point of the dummy head in case 4.

Figure 6.19: Absolute error per point in cases 3 and 4. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

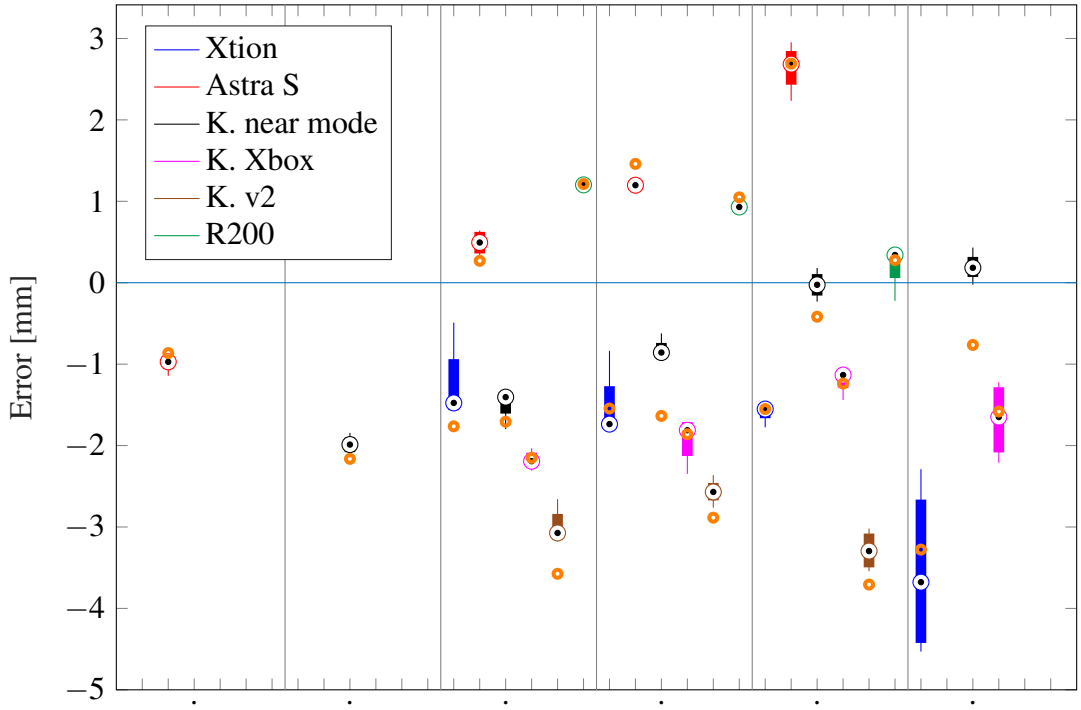


(a) Signed error per point of the dummy head in case 1.

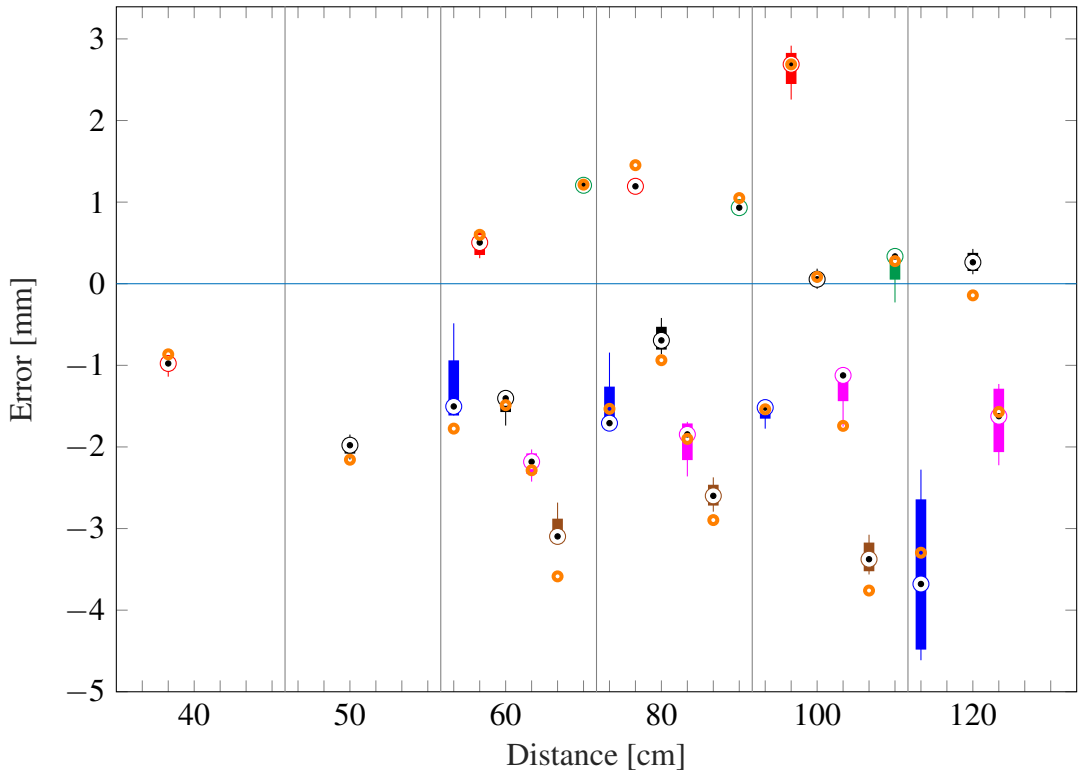


(b) Signed error per point of the dummy head in case 2.

Figure 6.20: Signed error per point in cases 1 and 2. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

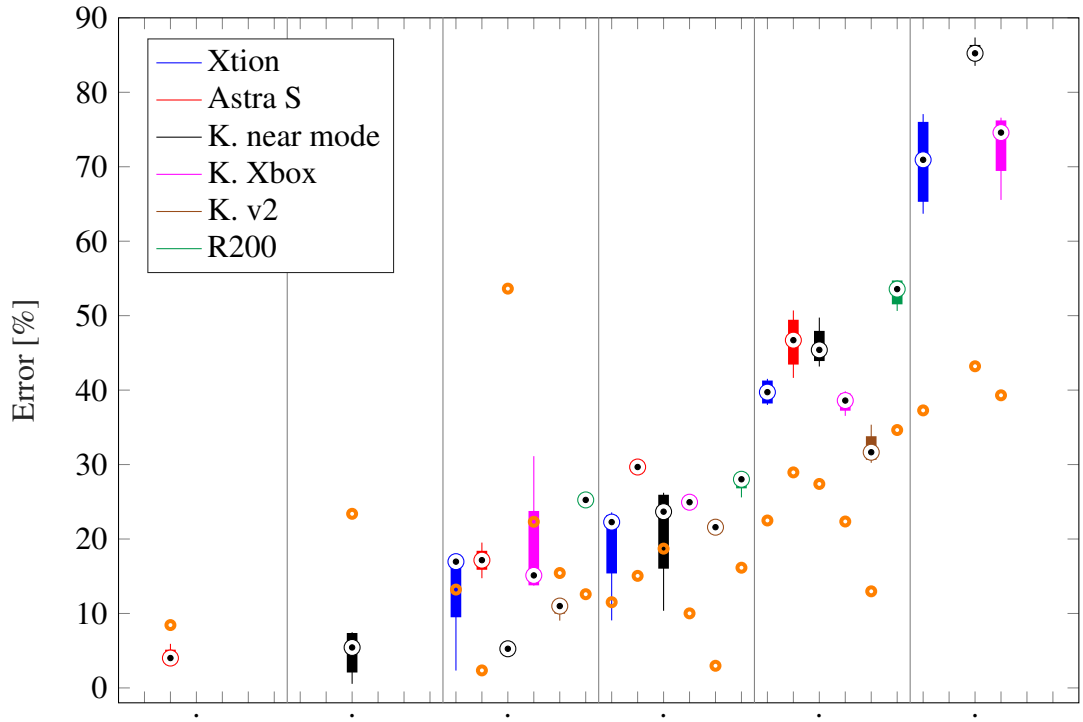


(a) Signed error per point of the dummy head in case 3.

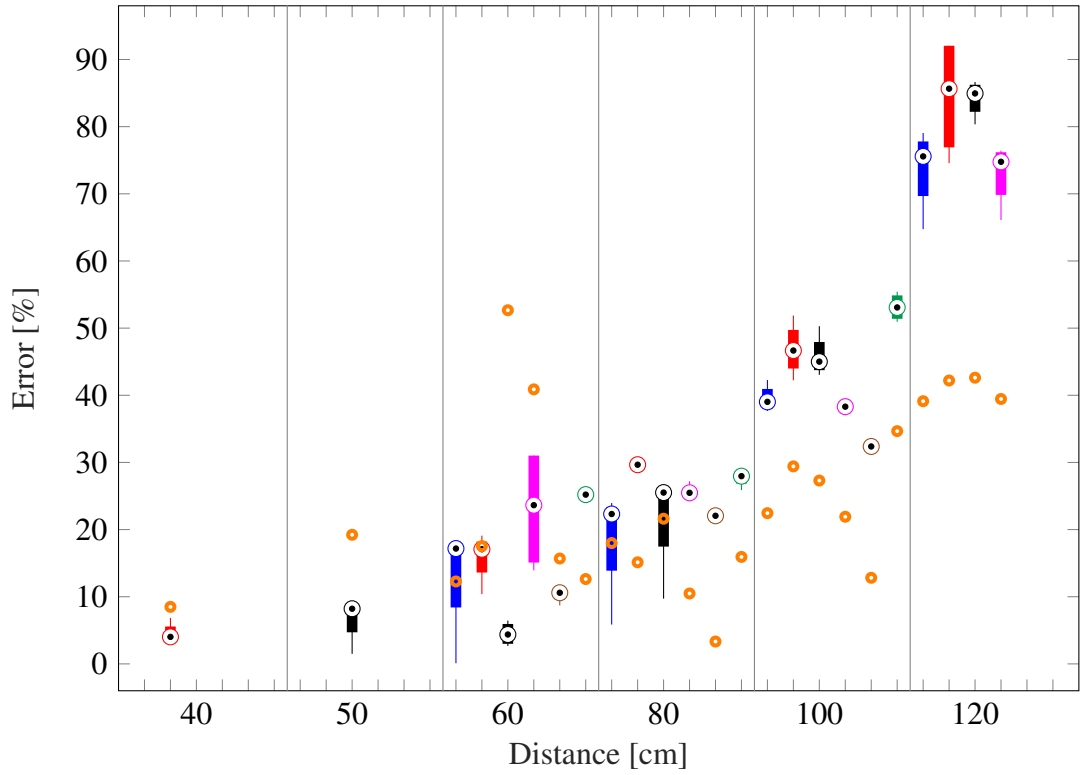


(b) Signed error per point of the dummy head in case 4.

Figure 6.21: Signed error per point in cases 3 and 4. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

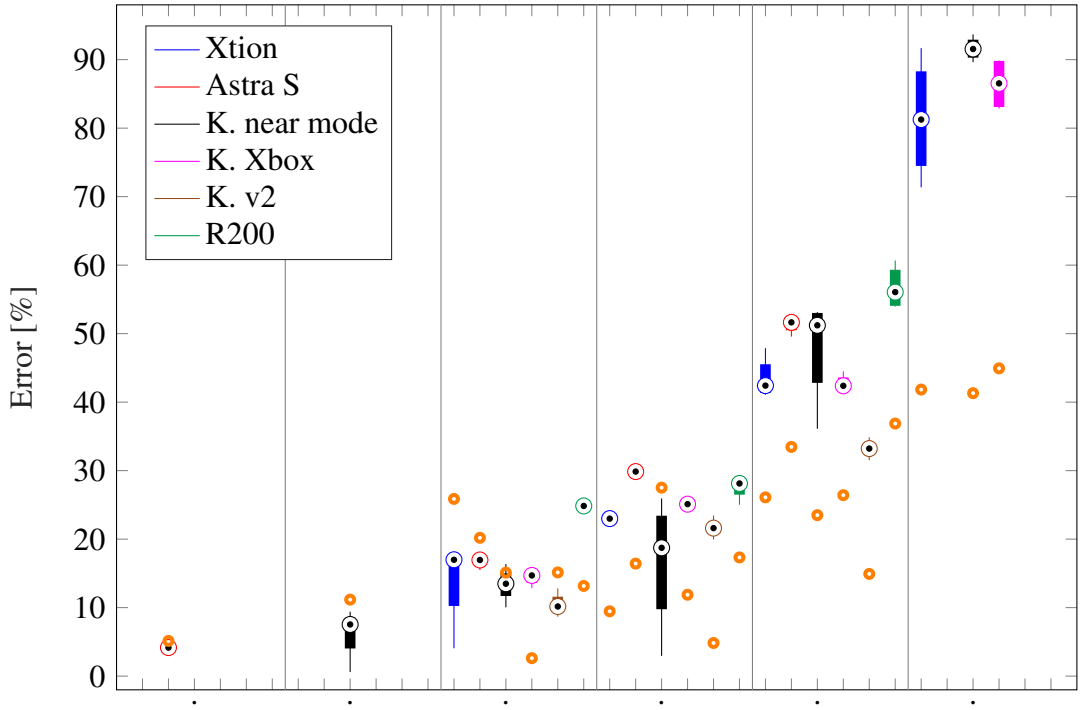


(a) Relative unsigned curvature error of the dummy head in case 1.

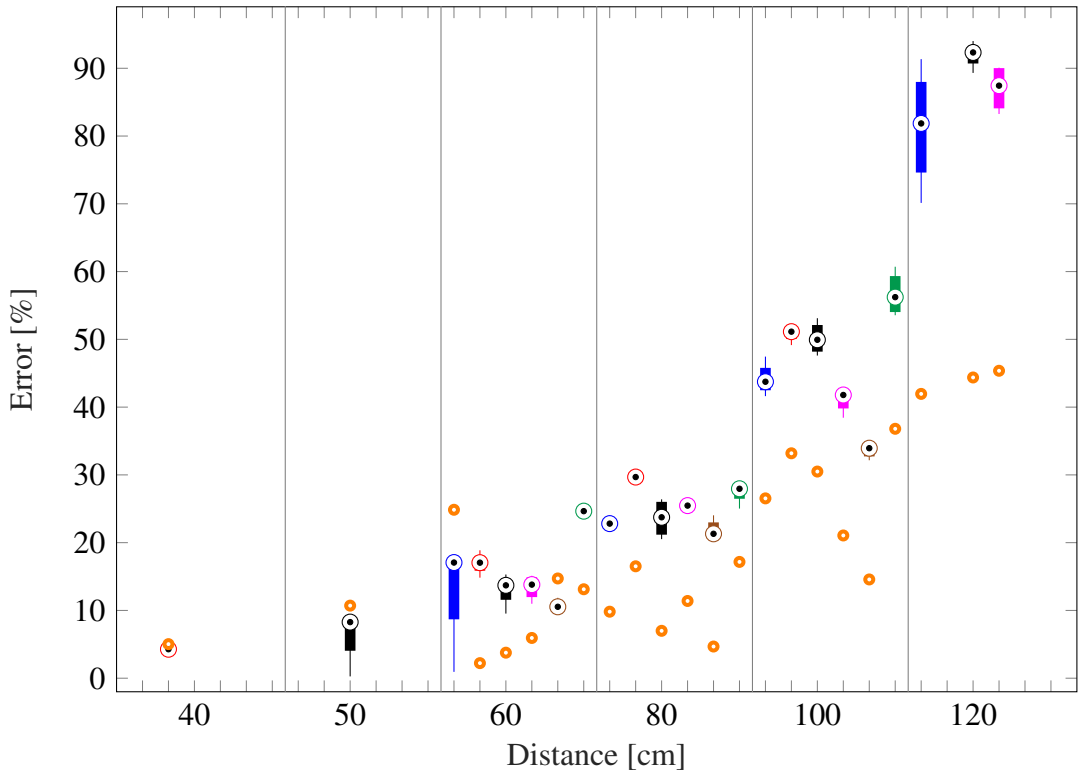


(b) Relative unsigned curvature error of the dummy head in case 2.

Figure 6.22: Relative unsigned curvature error in cases 1 and 2. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

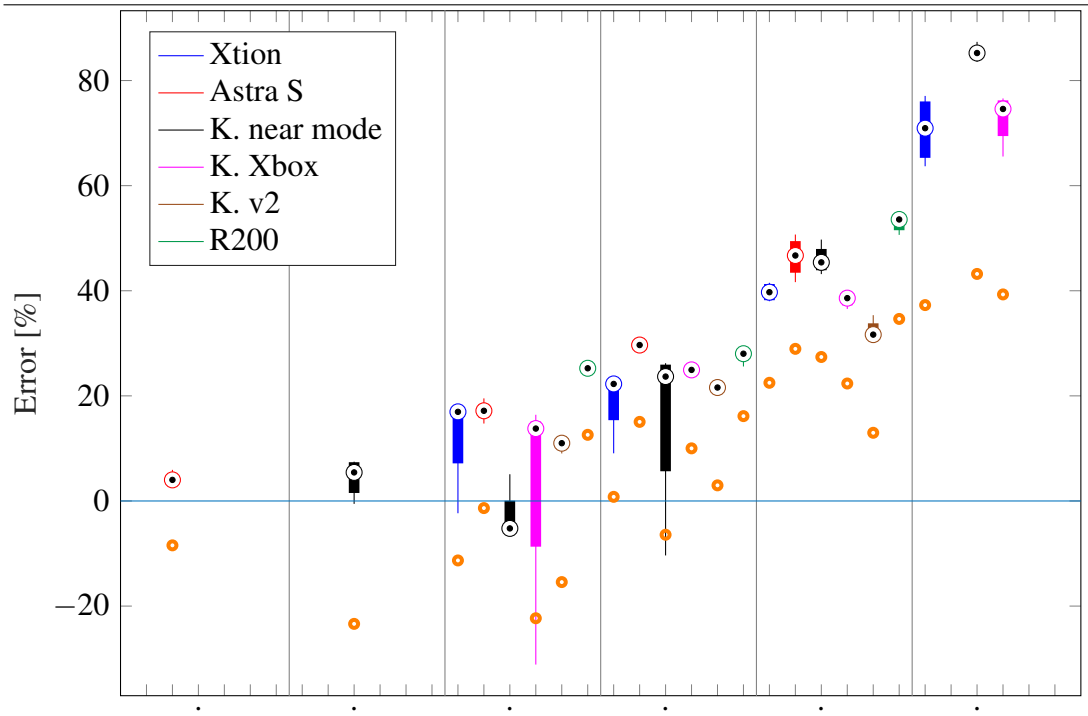


(a) Relative unsigned curvature error of the dummy head in case 3.

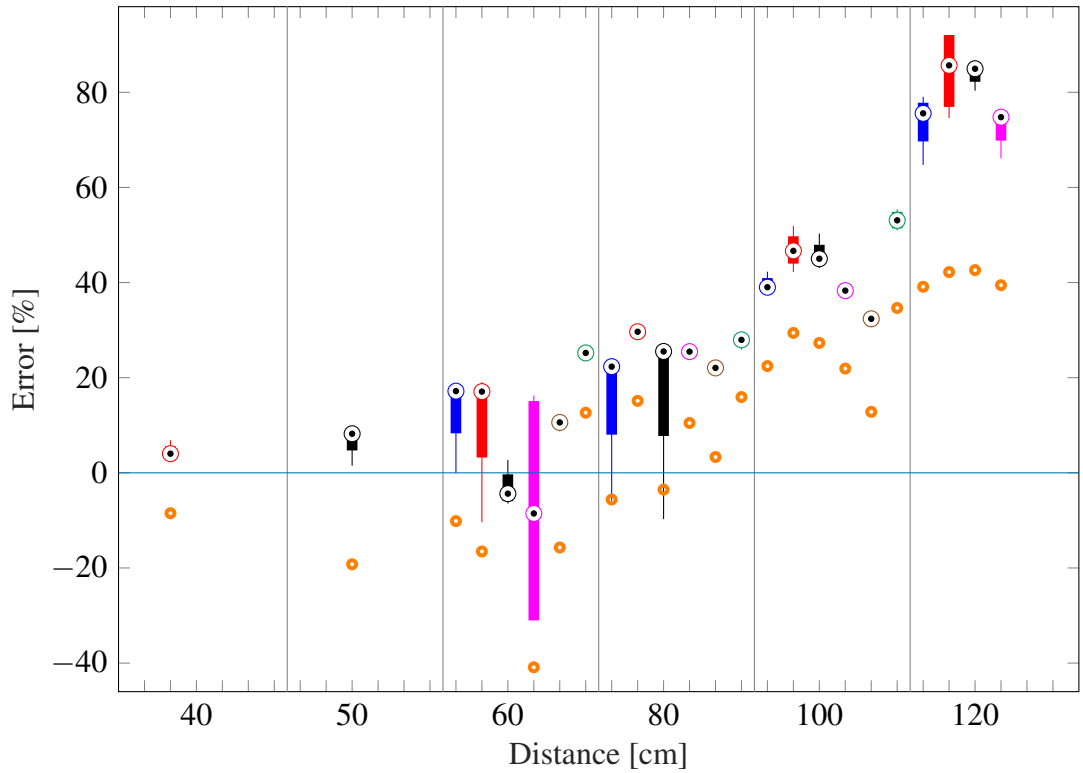


(b) Relative unsigned curvature error of the dummy head in case 4.

Figure 6.23: Relative unsigned curvature error per point in cases 3 and 4. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

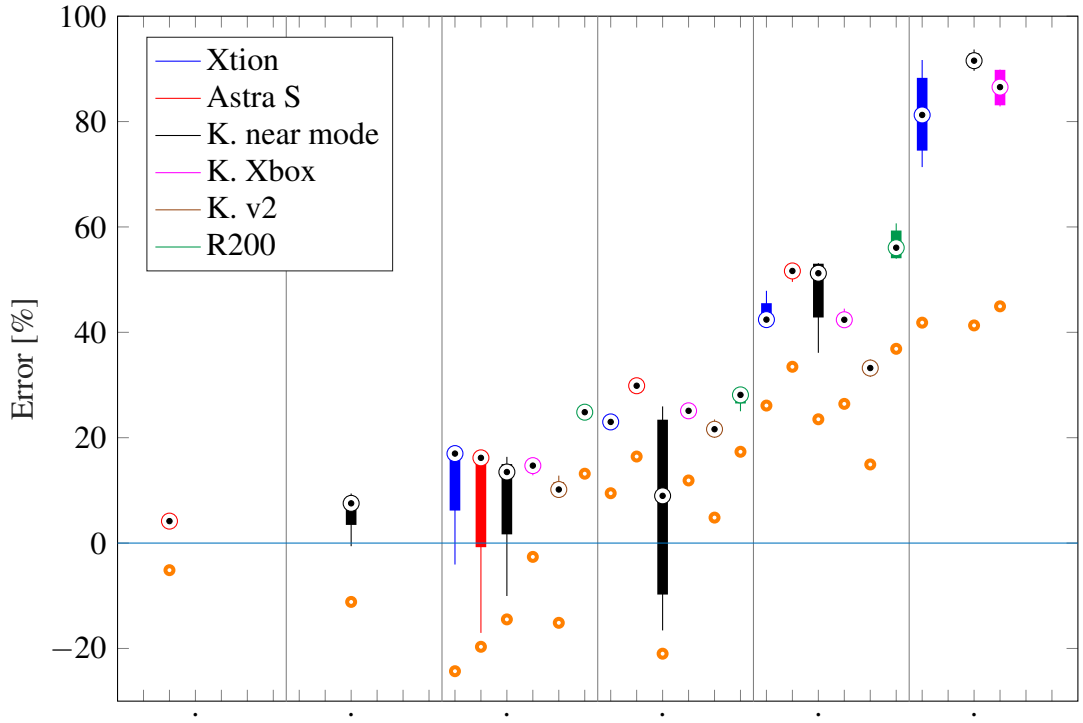


(a) Relative signed curvature error of the dummy head in case 1.

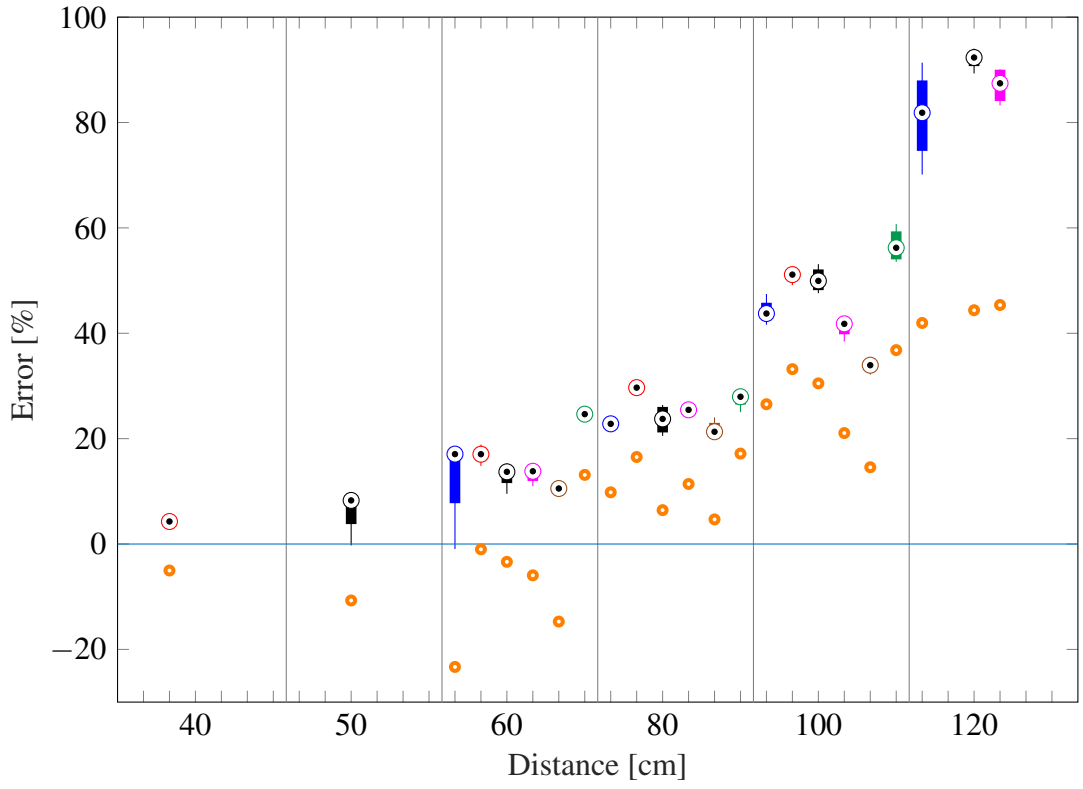


(b) Relative signed curvature error of the dummy head in case 2.

Figure 6.24: Relative signed curvature error in cases 1 and 2. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

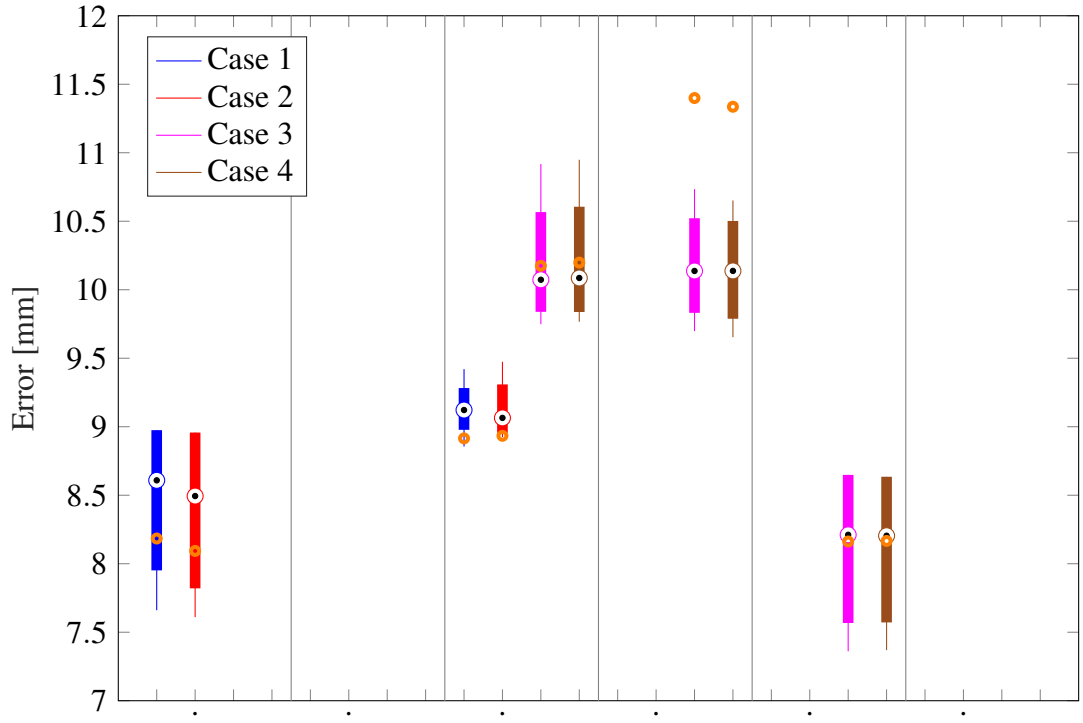


(a) Relative signed curvature error of the dummy head in case 3.

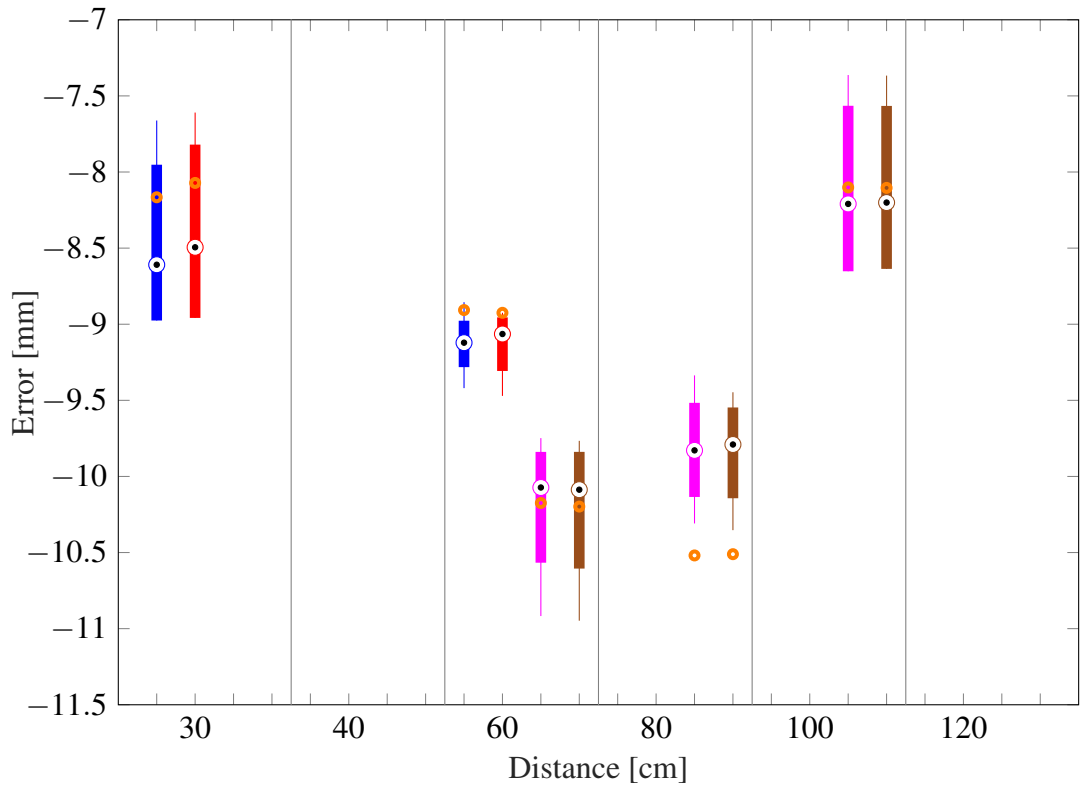


(b) Relative signed curvature error of the dummy head in case 4.

Figure 6.25: Relative signed curvature error per point in cases 3 and 4. The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

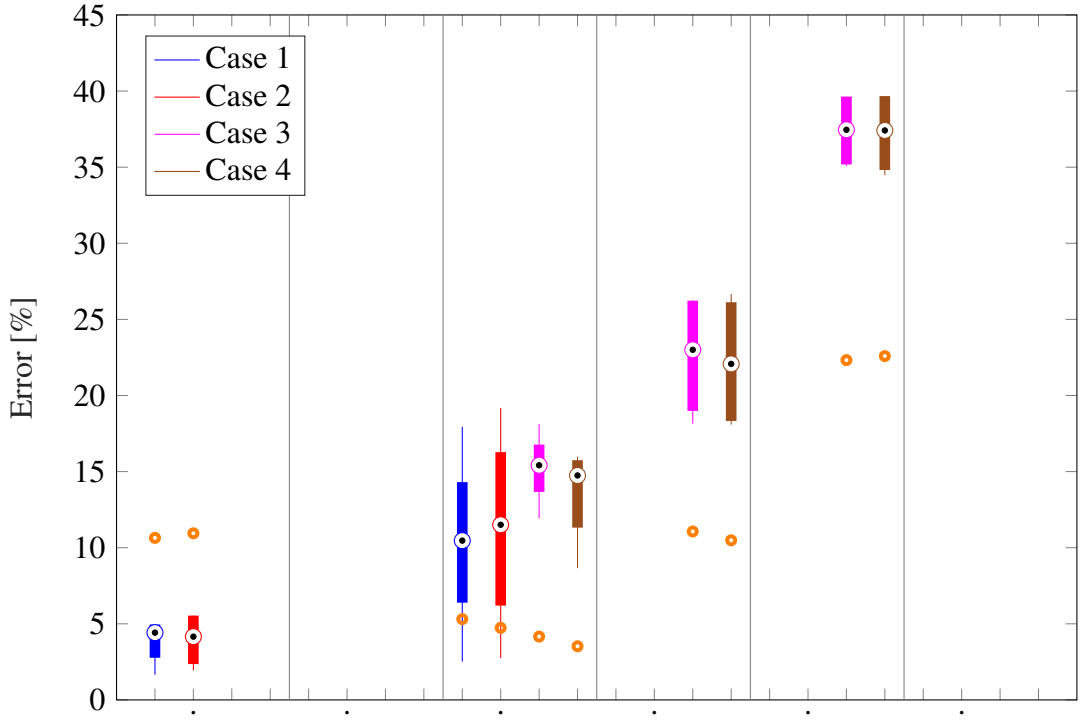


(a) Absolute error per point.

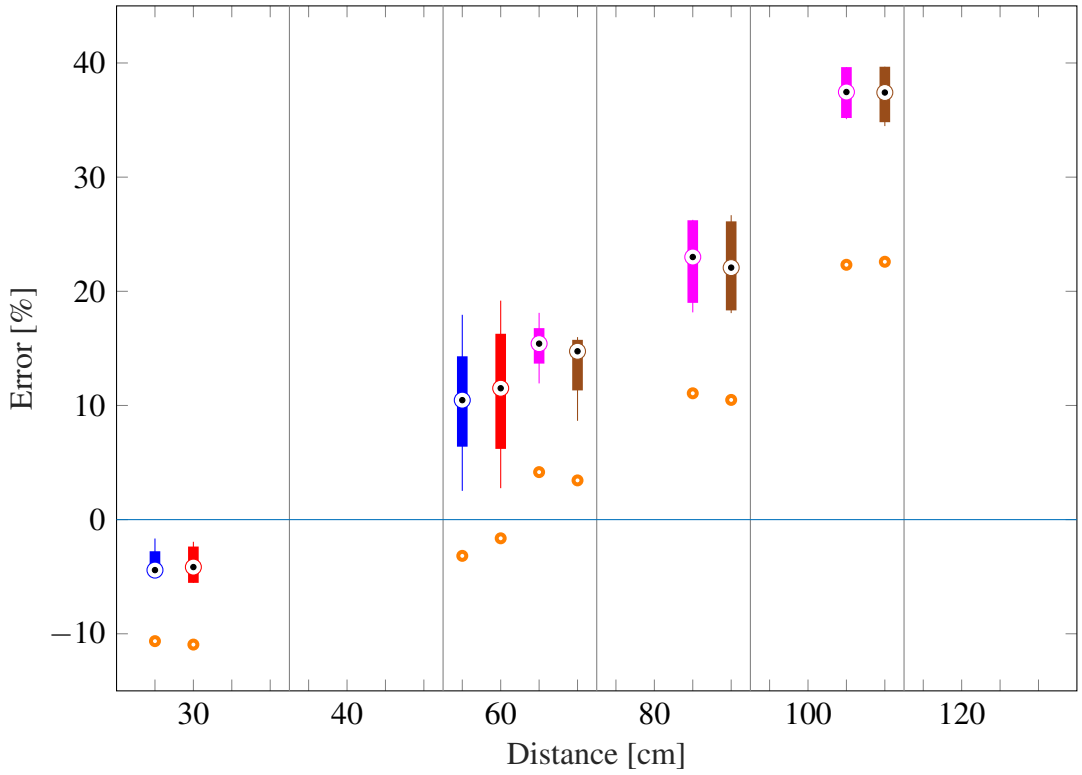


(b) Signed error per point.

Figure 6.26: Error per point given by the sensor D415 in all reconstruction cases. The cases are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.



(a) Relative unsigned curvature error.



(b) Relative signed curvature error.

Figure 6.27: Relative curvature error given by the sensor D415 in all reconstruction cases. The cases are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In this manuscript, an algorithm to automatically estimate head and ear measurements from uncolored 3D point clouds was presented. The feasibility of part of the approach was evaluated through a case study, where the estimated measurements were employed to assess the influence of anthropometric information on the interaural time difference (ITD) computation. Additionally, a technically-grounded guideline for RGBD sensor selection when reconstructing with KinFu was described. From such a guideline, a dataset with 3D models of four objects generated with six well-known sensors was created and is publicly available.

The automatically estimated head and ear measurements were determined by the dimensions required to personalize Head Related Transfer Functions (a person-dependent function to locate sound sources), they are:

- head width,
- head depth,
- head height,
- closest distance from the front of the head to the ear pits,
- closest distance from the back of the head to the ear pits, and
- helix height.

In order to estimate such distances, feature points were localized on 3D point clouds obtained with the RGBD sensor Asus Xtion Pro Live and the software KinFu. Such feature points are the ear pits, the top and back of the head, the chin, the forehead, the nose bridge and the nape. The algorithm was evaluated using 80 point clouds sets (each set consists of one 360° reconstruction, one right and one left profile point cloud) of 20 subjects generated with four different configurations: 1. the subject rotated in clockwise direction, and the reconstruction was performed at ≈ 5 fps, 2. the subject rotated in

counter-clockwise direction, and the reconstruction was performed at $\approx$ 5 fps, 3. the subject rotated in clockwise direction, and the reconstruction was performed at $\approx$ 10 fps, and 4. the subject rotated in counter-clockwise direction, and the reconstruction was performed at $\approx$ 10 fps. Experimental results showed an ear pit localization accuracy of 94.3 %. Moreover, an analysis of the measurements accuracy evidenced the existence of systematic (offset) errors on the dimensions measured along the horizontal axes, such as the head depth and width. After the systematic error correction, the median accuracy error ranged between 2.1 mm to 13.3 mm, meanwhile the mean ranged from 2.7 mm to 12.2 mm. The smallest mean and median values were exhibited by the helix height estimation; the largest median and mean errors were obtained when locating the nose bridge and the nape, respectively. The reproducibility of the algorithm was evaluated as the standard deviation of the estimated measurements of each subject's four models (each one was generated with a different configuration -reconstruction rate and rotation direction-). The median and mean standard deviations varied from 1.9 mm to 4.1 mm and from 2.2 mm to 5.0 mm, respectively. These results indicated that the presented approach can be used with different graphic cards and without restrictions of turning direction.

The feasibility of the algorithm to estimate the head measurements was evaluated through a case study, where the computed measurements were used to assess the influence of anthropometric data on the interaural time difference estimation. The evaluation was performed using 3D point clouds of 14 subjects. The mean accuracy error of the measurements range from 3.0 mm to 5.0 mm. These results are below the theoretical thresholds for ITD calculations described as the maximum allowed head measurements error. The mean values of the head dimensions were used to calculate the ITD; the results described that only very sensitive listeners will perceive a very small shift of the presented sound source position. Therefore, the automatic approach used to estimate the head measurements is a suitable method to personalize ITDs (component of the HRTF).

Since the combination of an RGBD sensor with KinFu represents a very promising alternative in the field of 3D reconstructions and nowadays there exist so many available sensors, a technically-grounded guideline for RGBD sensor selection when reconstructing with KinFu was created. The experiments that belong to this guideline were designed to imitate scenarios that researchers and end users might find, such as

1. low (8 fps) and full (30 fps) reconstruction rates since they depend on the graphic card's capability,
2. various distances between the sensor and the object to be reconstructed (from 40 cm to 120 cm depending on the sensor and the object),
3. uneven amount of input point clouds,
4. low number of input point clouds (800 point clouds).

Six well-known RGBD sensors were used to capture point clouds employed to reconstruct four different objects. The sensors are: a Microsoft Kinect Xbox 360, a Microsoft

Kinect in near mode, an Asus Xtion Pro Live, an Orbbec Astra S, and an Intel RealSense R200. The reconstructed objects are: a dummy head, a soccer ball, an American football, and a large rubber duck. Quantitative and qualitative evaluations were performed. The results showed that: 1. the outcome of Kinfu depends on the sensor used to record its input, and 2. the selection of a specific RGBD sensor depends on the requirements of each application. Nonetheless, the object to be scanned should be rich on non-uniform features and placed at a distance up to 100 cm to preserve details. A large influence of the reconstruction rate and the amount of input point clouds was not found. The median accuracy error was below 5.0 mm for all the sensors at all distances, which is an acceptable error for most applications. Regarding the mean curvature, the larger the distance, the larger the curvature loss. Moreover, the number of points of the reconstructions as point clouds is independent of the sensor technology and follows a decreasing quadratic function, as expected. Considering the comparable distance between the object and the sensor (from 60 cm to 120 cm), the Xtion Pro Live and the Astra S are recommended to work at 60 cm; at 80 cm the Xtion Pro Live surpassed the Astra S, conserving more details. The Kinect Xbox 360 exhibited average results; its lowest accuracy errors were at 100 cm. The sensors Kinect near mode and RealSense R200 were discarded due to the artifacts presented on the reconstructions.

In addition, a preliminary study was performed with one of the recently launched RGBD sensors: the Intel RealSense D415. In this study, only the dummy head was reconstructed. Horizontal artifacts were observed on the models, which might have been generated by the interaction of the projected laser with the object. Additionally, the number of input point clouds influenced the object's reconstruction: using all the point clouds recorded in one revolution, the dummy head was mostly reconstructed at 30 cm and 60 cm; meanwhile, using only 800 input point clouds the object was mostly reconstructed at 60 cm, 80 cm, and 100 cm. The median accuracy errors given by this sensor were at least 3.3 mm larger than the other sensors. Conversely, the curvature error presented a similar behavior than the other sensors.

7.2 Future Work

The head and ear measurements still present challenges due to the hair, which is part of the 3D reconstructions. Therefore, an alternative approach to generate a personalized head model from facial features, such as deformation of an a-priori model, is planned. Additionally, the detection of more feature points and the calculation of more measurements is intended, such as torso width and depth. Furthermore, since nowadays there exist many RGBD sensors, the estimation of such measurements from 3D point clouds generated with other sensors is also desired.

Keeping the guideline for RGBD sensor selection and the 3D models dataset up to date is intended; therefore, to include all the evaluations of the Intel RealSense D415 is a future work. Additionally, more sensors are planned to be evaluated, such as the Intel

Realsense D435 and future generations. The inclusion of more objects and scenes to reconstruct is also a valuable addition to both the guideline and the dataset. Moreover, a study about the influence of the remaining support cylinder in the reconstruction of the American football and the soccer ball is desired since these extra points may add features to the scene. In addition, further enhancement of the study with a deep examination of the effects of the number of input point clouds in cases 1 and 2, as well as with an evaluation of the mean opinion score [100] is contemplated. Furthermore, we plan to recalibrate the Kinect v2 and to try another driver such as [83] since in the literature [43, 55] the error of the Kinect v2 is lower than (or at least similar to) the other Kinects.

Appendix A

Accuracy per Point Evaluation of the Rubber Duck

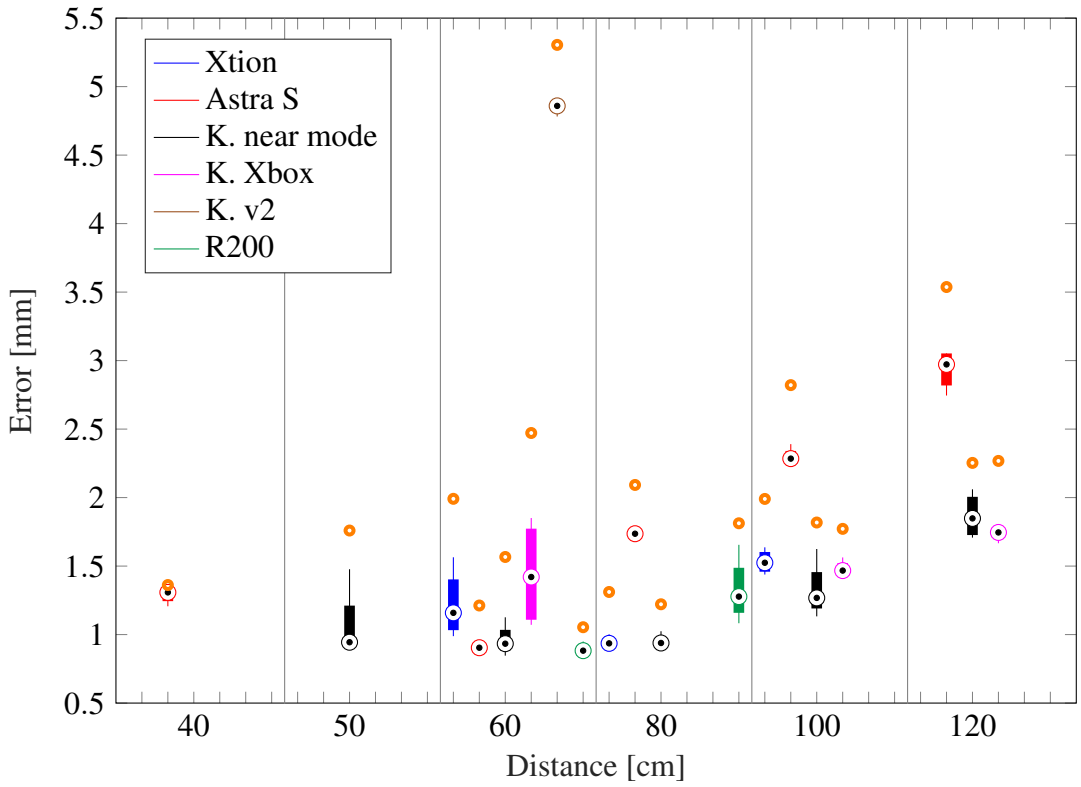


Figure A.1: Absolute error per point of the rubber duck in case 1 . The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

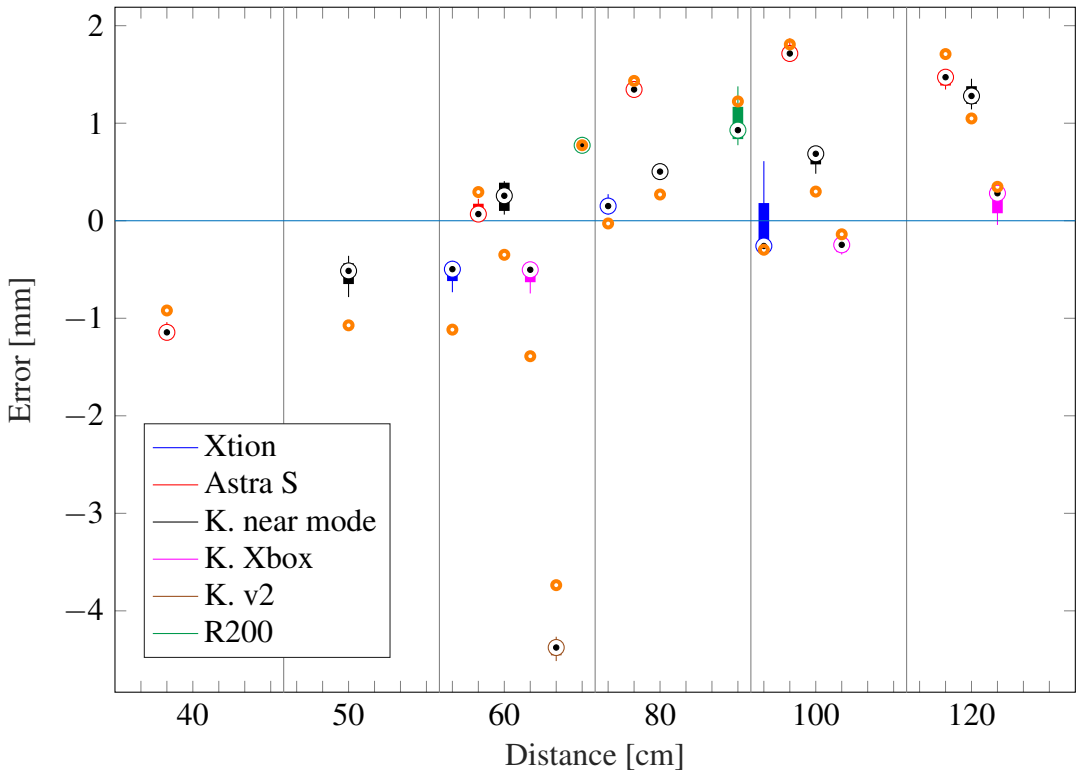


Figure A.2: Signed error per point of the rubber duck in case 1 . The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

Appendix B

Curvature Accuracy Evaluation of the Rubber Duck

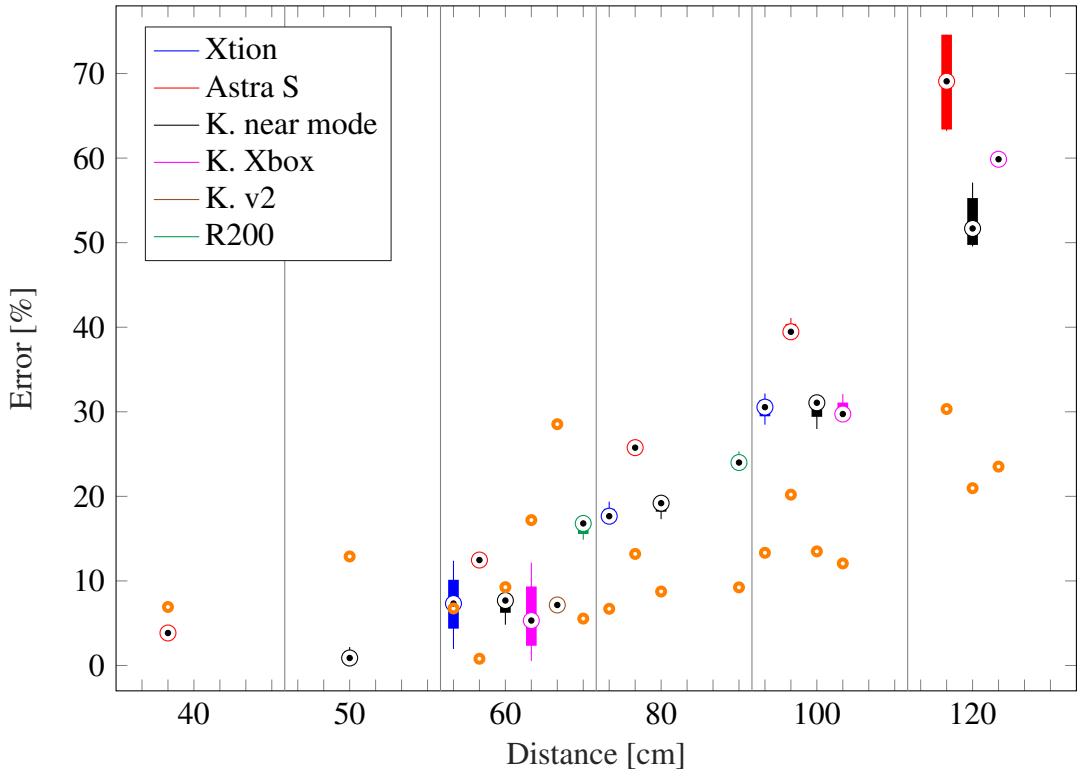


Figure B.1: Relative unsigned curvature error of the rubber duck in case 1 . The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

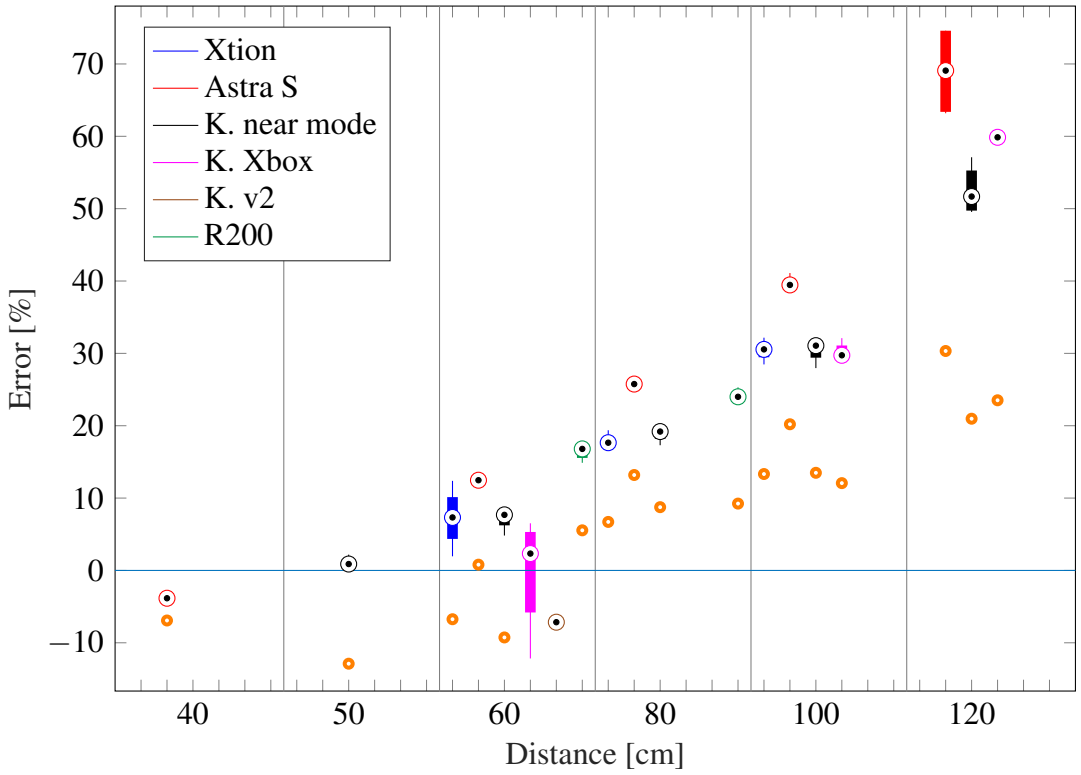


Figure B.2: Relative signed curvature error of the rubber duck in case 1 . The sensors are organized at each distance as is shown in the legend. The top- and bottom most limits of the lines are the maximum and minimum values, respectively. The upper and lower boxes' lines are the third and first quartile, respectively. The black dots give the median and the orange rings the mean.

Symbols

φ	Azimuth angle on the horizontal plane
θ_j	Angle between a interest point and a neighbor j
a	Averaged head radius
b	Distance from the ear pits to the nape
$\vec{C}$	Camera center
c_0	Speed of sound
d	Head depth measured between the forehead and the back of the head
d_c	Head depth calculated as $0.5\,(f_h + b)$
d_{x_2}	Head height from the CIPIC database
d_{x_3}	Head depth from the CIPIC database
d_{x_4}	Vertical pinna offset from the CIPIC database
d_{x_5}	Horizontal pinna offset from the CIPIC database
e_c	Accuracy error of the head measurements in the case study $(m_t - m_e)$
e_m	Accuracy error of the head measurements $(m_e - m_t)$
e_p	Accuracy error per point
e_v	Curvature accuracy error
f	Focal length
f_h	Distance from the ear pits to the nose bridge
g	Ground-truth values for the RGBD sensors study
GR	Good reconstructions
h_c	Head height calculated from the top to the center of the head
h_e	Helix height
h_h	Head height measured between the top of the head and the chin
k	Wave number
K	Intrinsic camera parameters
K_n	Neighborhood
m_e	Estimated head measurement
m_t	Ground-truth head measurement
$\vec{n}_{p_i}$	Normal of the interest point p_i

$\vec{p}$	Principal point
p_i	Interest point
p_j	The $j-th$ neighbor of p_i
p_x	Principal point offset along the x axis
p_y	Principal point offset along the y axis
R	Relative camera rotation to the world
r	Radius
s	Reconstructions' values
SR	Success rate of
t	Relative camera translation to the world
TR	Total number of reconstructions per sensor-distance
w	Head width measured between the tragi
w_c	Head width measured between the center of the head and a tragus
$\vec{x}$	Projected point in the image plane
$\vec{X}$	Point in the three-dimensional world
x	Axis or coordinate
y	Axis or coordinate
z	Axis or coordinate
X	Coordinate of the point $\vec{X}$
Y	Coordinate of the point $\vec{X}$
Z	Coordinate of the point $\vec{X}$

Abbreviations

2D	Two dimensional
3D	Three dimensional
AS	Orbbec Astra S
CUDA	Compute unified device architecture
DOF	Degree of freedom
fps	Frames per second
GPU	Graphics processing unit
HRTF	Head related transfer function
ICP	Iterative closest point
IR	Infrared
ITD	Interaural time difference
JND	Just noticeable difference
K2	Microsoft Kinect v2
KN	Microsoft Kinect in near mode
KX	Microsoft Kinect Xbox 360
MRI	Magnetic resonance image
PCA	Principal component analysis
PCL	Point cloud library
R2	Intel RealSense R200
RGB	Red, green, and blue
RGBD	Red, green, blue, and depth
RMS	Root mean square
ROI	Region of interest
SfM	Structure from motion
ToF	Time of flight
TSDF	Truncated signed distance function
vpa	Voxels per axis
XT	Asus Xtion Pro Live

Bibliography

[1] Artec Eva 3D Scanner. https://www.artec3d.com/portable-3d-scanners/artec-eva. (Visited on Nov 23rd, 2017) [Online].

[2] CloudCompare. http://cloudcompare.org/. (Visited on March 1st, 2018) [Online].

[3] Intel RealSense D415 Specifications. https://click.intel.com/intelr-realsensetm-depth-camera-d415.html. (Visited on Dec 18th, 2018) [Online].

[4] KinFu. https://github.com/PointCloudLibrary/pcl/tree/master/gpu/kinfu. (Visited on Nov 23rd, 2017) [Online].

[5] Microsoft Kinect for Windows. https://www.cnet.com/products/microsoft-kinect-for-windows/specs/. (Visited on Nov 14th, 2018) [Online].

[6] Near Mode: What it is (and isn't). https://blogs.msdn.microsoft.com/kinectforwindows/2012/01/20/near-mode-what-it-is-and-isnt/. (Visited on Nov 14th, 2018) [Online].

[7] Orbbec Astra, Astra S & Astra Pro. https://orbbec3d.com/product-astra/. (Visited on Nov 13th, 2018) [Online].

[8] Point Cloud Library (PCL). http://pointclouds.org/. (Visited on April 12th, 2018) [Online].

[9] RealSense. http://wiki.ros.org/RealSense. (Visited on May 12th, 2017) [Online].

[10] ReconstructMe. http://reconstructme.net/. (Visited on Dec 25th, 2018) [Online].

[11] Report: Here Are Kinect's Technical Specs. https://kotaku.com/5576002/here-are-kinects-technical-specs. (Visited on Nov 14th, 2018) [Online].

[12] Revealing Kinect for Windows v2 Hardware. https://blogs.msdn.microsoft.com/kinectforwindows/2014/03/27/revealing-kinect-for-windows-v2-hardware/. (Visited on Nov 16th, 2018) [Online].

[13] ROS. http://www.ros.org/. (Visited on April 12th, 2018) [Online].

[14] Time-of-Flight Principle. https://www.terabee.com/time-of-flight-principle/. (Visited on Nov 15th, 2018) [Online].

[15] Xtion PRO LIVE. https://www.asus.com/de/3D-Sensor/Xtion_PRO_LIVE/specifications/. (Visited on Nov 13th, 2018) [Online].

[16] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building Rome in a Day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79. Ieee, sep 2009.

[17] A. Al-Nuaimi, M. Piccolrovazzi, S. Gedikli, E. Steinbach, and G. Schroth. Indoor Location Retrieval using Shape Matching of KinectFusion Scans to Large-Scale Indoor Point Clouds. In *Eurographics Workshop on 3D Object Retrieval*, 2015.

[18] V. Algazi, R. Duda, D. Thompson, and C. Avendano. The CIPIC HRTF Database (website). https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/. (Visited on April 16th, 2018) [Online].

[19] V. Algazi, R. Duda, D. Thompson, and C. Avendano. The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, number October, pages 99–102, 2001.

[20] V. R. Algazi, C. Avendano, and R. O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *Journal of the Audio Engineering Society*, 49(6):472–479, 2001.

[21] S. T. Barnard and M. A. Fischler. Computational Stereo. *ACM Computing Surveys*, 14(4):553–572, 1982.

[22] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded Up Robust Features. In *European conference on computer vision*, 2006.

[23] G. Birbilis. Kinect for Xbox 360 and Kinect for Windows (KfW) Specs. https://zoomicon.wordpress.com/2015/07/28/kinect-for-xbox-360-and-kinect-for-windows-kfw-v1-specs/. (Visited on Dec 21st, 2018) [Online].

[24] G. Böer, F. Hahmann, I. Buhr, H. Essig, and H. Schramm. Detection of Facial Landmarks in 3D Face Scans Using the Discriminative Generalized Hough Transform (DGHT). In *Portability of TV-Regularized Reconstruction Parameters to Varying Data Sets*, pages 299–304. Springer Vieweg, 2015.

[25] R. Bomhardt, M. de la Fuente Klein, and J. Fels. A High-Resolution Head-Related Transfer Function and Three-Dimensional Ear Model Database. In *172nd Meeting of the Acoustical Society of America*, volume 050002, page 050002, 2016.

[26] R. Bomhardt, I. C. P. Mejia, A. Zell, and J. Fels. Required Measurement Accuracy of Head Dimensions for Modeling the Interaural Time Difference. *Journal of the Audio Engineering Society*, 66(3):114–126, 2018.

[27] C. Bregler, A. Hertzmann, and H. Biermann. Recovering Non-Rigid 3D Shape from Image Streams. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 17, pages 706–713, 2000.

[28] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[29] M. Bueno, L. Díaz-Vilariño, J. Martínez-Sánchez, H. González-Jorge, H. Lorenzo, and P. Arias. Metrological Evaluation of KinectFusion and its Comparison with Microsoft Kinect Sensor. *Measurement: Journal of the International Measurement Confederation*, 73:137–145, 2015.

[30] Cannadian Association of Optometrists. Binocular Vision. https://opto.ca/health-library/binocular-vision, 2018. (Visited on Nov 17th, 2018) [Online].

[31] S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun. A Large Dataset of Object Scans. 2016.

[32] C. Collen. Introducing the Intel® RealSense™ R200 Camera (world facing). https://software.intel.com/en-us/articles/realsense-r200-camera, 2015. (Visited on Dec 27th, 2018) [Online].

[33] E. Crognier. Climate and Anthropometric Variations in Europe and the MediterraneanArea. *Annals of Human Biology*, 8(2):99–107, 1981.

[34] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure From Motion Without Correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, volume 2, pages 557–564.

[35] M. Dinakaran, P. Grosche, F. Brinkmann, and Weinzierl. Extraction of Anthropometric Measures from 3D-Meshes for the Individualization of Head-Related Transfer Functions. In Audio Engineering Society, editor, *140th Audio Engineering Society Convention*, 2016.

[36] Z. Ding, L. Zhang, and H. Li. A Novel 3D Ear Identification Approach Based on Sparse Representation. In *2013 20th IEEE International Conference on Image Processing*, pages 4166–4170, Melbourne, Australia, 2013. IEEE.

[37] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision*, 101(3):437–458, 2013.

[38] L. G. Farkas, J. C. Posnick, and T. M. Hreczko. Anthropometric Growth Study of the Head. *Cleft Palate-Craniofacial Journal*, 29(4):303–307, 1992.

[39] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth Mapping Using Projected Patterns, 2012.

[40] R. Garg, A. Roussos, and L. Agapito. Dense Variational Reconstruction of Non-Rigid Surfaces from Monocular Video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013.

[41] A. Geiger, M. Roser, and R. Urtasun. Efficient Large-Scale Stereo Matching. *Computer Vision – ACCV 2010*, 2010.

[42] A. Geiger, J. Ziegler, and C. Stiller. StereoScan: Dense 3D Reconstruction in Real-Time. In *IEEE Intelligent Vehicles Symposium*, number Iv, pages 963–968, 2011.

[43] H. Gonzalez-Jorge, P. Rodríguez-Gonzálvez, J. Martínez-Sánchez, D. González-Aguilera, P. Arias, M. Gesto, and L. Díaz-Vilariño. Metrological Comparison Between Kinect I and Kinect II Sensors. *Measurement: Journal of the International Measurement Confederation*, 70:21–26, 2015.

[44] F. Groh, B. Resch, and H. P. Lensch. Multi-View Continuous Structured Light Scanning. In *German Conference on Pattern Recognition*, volume 10496 LNCS, pages 377–388, 2017.

[45] K. Häming and G. Peters. The Structure-from-Motion Reconstruction Pipeline - A Survey with Focus on Short Image Sequences. *Kybernetika*, 46(5):926–937, 2010.

[46] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2003.

[47] C. Hoene, I. C. Patino Mejia, and A. Cacerovschi. MySofa—Design Your Personal HRTF. *142nd Convention of the Audio Engineering Society*, 2017.

[48] S. Izadi, A. Davison, A. Fitzgibbon, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, and D. Freeman. Kinect Fusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *24th annual ACM symposium on User interface software and technology - UIST '11*, page 559, 2011.

[49] T. Jebara, A. Azarbayejani, and A. Pentland. 3D Structure from 2D Motion. *IEEE Signal Processing Magazine*, 16(3):66–84, 1999.

[50] B. F. G. Katz and G. Parseihian. Perceptually Based Head-Related Transfer Function Database Optimization. *The Journal of the Acoustical Society of America*, 131(2):EL99–EL105, 2012.

[51] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik. Intel RealSense Stereoscopic Depth Cameras. In *IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition Workshops, volume 2017-July, pages 1267–1276, 2017.

[52] J. J. Koenderink and A. J. van Doorn. Affine Structure from Motion. *Journal of the Optical Society of America*, 8(2):377, 1991.

[53] M. Korn and J. Pauli. KinFu MOT : KinectFusion with Moving Objects Tracking. In *VISAPP (3)*, pages 648–657, 2015.

[54] G. F. Kuhn. Model for the Interaural Time Differences in the Azimuthal Plane. *The Journal of the Acoustical Society of America*, 62(1975):157, 1977.

[55] E. Lachat, H. Macher, M. A. Mittet, T. Landes, and P. Grussenmeyer. First Experiences with Kinect V2 Sensor for Close Range 3D Modelling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, (XL-5/W4), 2015.

[56] R. M. Laing, E. J. Holland, C. A. Wilson, and B. E. Niven. Development of Sizing Systems for Protective Clothing for the Adult Male. *Ergonomics*, 42(10):1249–1257, 1999.

[57] R. Lange. *3D Time-of-Flight Distance Measurement with Custom Solid-State Image Sensors in CMOS/CCD-Technology*. PhD thesis, Universität-Gesamthochschule Siegen, 2000.

[58] J. Lei, X. You, M. Abdel-mottaleb, S. Member, and M. Abdel-mottaleb. Automatic Ear Landmark Localization, Segmentation, and Pose Classification in Range Images. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(2):165–176, 2016.

[59] Y. Liu, X. Peng, W. Zhou, B. Liu, and A. Gerndt. Template-Based 3D Reconstruction of Non-rigid Deformable Object from Monocular Video. *3D Research*, 9:1–12, 2018.

[60] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[61] M. C. F. Macedo, A. L. J. Apolinário, and A. C. S. Souza. KinectFusion for Faces: Real-Time 3D Face Tracking and Modeling Using a Kinect Camera for a Markerless AR System. *SBC Journal on 3D Interactive Systems*, 4(2):2–7, 2013.

[62] E. A. Macpherson and J. C. Middlebrooks. Listener Weighting of Cues for Lateral Angle: the Duplex Theory of Sound Localization Revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, 2002.

[63] K. McMullen, A. Roginska, and G. Wakefield. Subjective Selection of Head-Related Transfer Functions (HRTF) Based on Spectral Coloration and Interaural Time Differences (ITD) Cues. In Audio Engineering Society, editor, *133rd Audio Engineering Society Convention*, 2012.

[64] S. Meister, P. Kohli, S. Izadi, M. Hämmerle, C. Rother, and D. Kondermann. When can we use KinectFusion for ground truth acquisition? In *Workshop on Color-Depth Camera Fusion in Robotics*, pages 3–8, 2012.

[65] P. Mihelich. openni_launch. http://wiki.ros.org/openni_launch. (Visited on May 12th, 2018) [Online].

[66] P. Mihelich, S. Gedikli, and B. R. Radu. openni_camera. http://wiki.ros.org/openni_camera. (Visited on May 12th, 2017) [Online].

[67] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real Time Localization and 3D Reconstruction. In *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006.

[68] Nerian Vision Technologies. Karmin2 – Nerian's 3D Stereo Camera. https://nerian.com/products/karmin2-3d-stereo-camera/, 2018. (Visited on Nov 17th, 2018) [Online].

[69] R. A. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion : Reconstruction and Tracking of Non-rigid Scenes in Real-Time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 343–352, 2015.

[70] R. A. Newcombe, D. Molyneaux, D. Kim, A. J. Davison, J. Shotton, S. Hodges, A. Fitzgibbon, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-time Dense Surface Mapping and Tracking. In IEEE, editor, *10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011*, pages 127–136, 2011.

[71] C. A. Nock, O. Taugourdeau, S. Delagrange, and C. Messier. Assessing the Potential of Low-Cost 3D Cameras for the Rapid Measurement of Plant Woody Structure. *Sensors*, 13(12):16216–16233, 2013.

[72] O. Ozyesil, V. Voroninski, R. Basri, and A. Singer. A Survey of Structure from Motion. pages 1–40, 2017.

[73] G. R. P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, P. Cignoni, P. Cignoni, M. Callieri, M. Callieri, M. Corsini, M. Corsini, M. Dellepiane, M. Dellepiane, F. Ganovelli, F. Ganovelli, G. Ranzuglia, and G. Ranzuglia. MeshLab: an Open-Source Mesh Processing Tool. In *Sixth Eurographics Italian Chapter Conference*, pages 129–136, 2008.

[74] D. Pagliari, F. Menna, R. Roncella, F. Remondino, and L. Pinto. Kinect Fusion Improvement Using Depth Camera Calibration. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5(June):479–485, 2014.

[75] G. Parseihian and B. F. G. Katz. Rapid Head-Related Transfer Function Adaptation Using a Virtual Auditory Environment. *The Journal of the Acoustical Society of America*, 131(4):2948–2957, 2012.

[76] I. C. Patino Mejia and A. Zell. Head Measurements from 3D Point Clouds. In *6th I International Conference on Image Processing Theory, Tools and Applications (IPTA 2016)*, 2017.

[77] I. C. Patino Mejia and A. Zell. 3D Reconstructions with KinFu Using Different RGBD Sensors. In *3rd IEEE International Conference on Image Processing, Applications and Systems (IPAS 2018)*, 2018.

[78] A. Payne, A. Daniel, A. Mehta, B. Thompson, C. S. Bamji, D. Snow, H. Oshima, L. Prather, M. Fenton, L. Kordus, P. O'Connor, R. McCauley, S. Nayak, S. Acharya, S. Mehta, T. Elkhatib, T. Meyer, T. O'Dwyer, T. Perry, V. H. Chan, V. Wong, V. Mogallapu, W. Qian, and Z. Xu. A 512×424 CMOS 3D Time-of-Flight Image Sensor with Multi-Frequency Photo-Demodulation up to 130MHz and 2GS/s ADC, 2014.

[79] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris. 3D Facial Landmark Detection Under Large Yaw and Expression Variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2012.

[80] S. Prakash and P. Gupta. A Rotation and Scale Invariant Technique for Ear Detection in 3D. *Pattern Recognition Letters*, 33(14):1924–1931, 2012.

[81] L. Rayleigh. XII. On Our Perception of Sound Direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[82] B. Resch, H. P. A. Lensch, O. Wang, M. Pollefeys, and A. Sorkine-Hornung. Scalable Structure from Motion for Densely Sampled Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3936–3944, 2015.

[83] H. Richard. Kinect2. https://github.com/doge-of-the-day/kinect2. (Visited on March 6th, 2018) [Online].

[84] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, T. Brill, D. Hoeferlin, and D. Burnsides. CAESAR Final report, volume I: Summary. Technical report.

[85] K. M. Robinette, H. Daanen, and E. Paquet. The CAESAR project: a 3-D surface anthropometry survey. In *Second International Conference on 3D Digital Imaging and Modeling Cat NoPR00062*, pages 380–386, 1999.

[86] H. Roth and V. Marsette. Moving Volume KinectFusion. In *British Machine Vision Conference*, pages 112.1—-112.11, 2012.

[87] M. Rothbucher, T. Habigt, J. Habigt, T. Riedmaier, and K. Diepold. Measuring Anthropometric Data for HRTF Personalization. In *6th International Conference on Signal Image Technology and Internet Based Systems, SITIS*, pages 102–106, 2010.

[88] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB : an Efficient Alternative to SIFT or SURF. In *IEEE international conference on Computer Vision (ICCV)*, 2011.

[89] B. L. Schwartz and J. H. Krantz. Sensation and Perception. https://edge.sagepub.com/schwartz. (Visited on Dec 18th, 2018) [Online].

[90] M. P. Segundo, C. Queirolo, O. R. P. Bellon, and L. Silva. Automatic 3D Facial Segmentation and Landmark Detection. In *4th International conference on Image Analysis and Processing, ICIAP 2007*, number Iciap, pages 431–436, 2007.

[91] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. BigBIRD: A Large-Scale 3D Database of Object Instances. In *International Conference on Robotics and Automation (ICRA 2014)*, number 1, 2014.

[92] R. Smeenk. Kinect V1 and Kinect V2 Fields of View Compared. https://smeenk.com/kinect-field-of-view-comparison/. (Visited on Dec 27th, 2018) [Online].

[93] A. Smith. Official Xbox One and Kinect 2 Dimensions Revealed. https://www.gamepur.com/news/12519-official-xbox-one-and-kinect-2-dimensions-revealed.html. (Visited on Dec 25th, 2018) [Online].

[94] J. Sturm, E. Bylow, F. Kahl, and D. Cremers. CopyMe3D: Scanning and Printing Persons in 3D. In *German Conference on Pattern Recognition (GCPR)*, pages 405–414, 2013.

[95] M. Szymczyk. How Does The Kinect 2 Compare To The Kinect 1? https://zugara.com/how-does-the-kinect-2-compare-to-the-kinect-1. (Visited on Nov 14th, 2018) [Online].

[96] L. Tim. astra_camera. http://wiki.ros.org/astra_camera. (Visited on May 12th, 2017) [Online].

[97] E. A. Torres-Gallegos, F. Orduña-Bustamante, and F. Arámbula-Cosío. Personalization of Head-Related Transfer Functions (HRTF) Based on Automatic Photo-Anthropometry and Inference from a Database. *Applied Acoustics*, 97:84–95, 2015.

[98] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and Modeling Non-Rigid Objects with Rank Constraints. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages I–493–I–500, 2001.

[99] B. Triggs, P. Mclauchlan, R. Hartley, A. Fitzgibbon, B. Triggs, P. Mclauchlan, R. Hartley, A. Fitzgibbon, B. Ajustment, S. B. Triggs, A. Zisserman, R. Szeliski, V. Algorithms, B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon. Bundle Adjustment – A Modern Synthesis. In *International Workshop on Vision Algorithms*, 2000.

[100] I. T. Union. Methods for Objective and Subjective Assessment of Speech and Video Quality. Technical report, 2016.

[101] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D Body Scans from Noisy Image and Range Data. In *IEEE International Conference onComputer Vision (ICCV 2011)*, pages 1951–1958. Dept. of Computer Science, Brown University, Ieee, 2011.

[102] T. Wiedemeyer. IAI Kinect2. https://github.com/code-iai/iai_kinect2.

[103] R. S. Woodworth. Experimental Psychology. *The Journal of Nervous and Mental Disease*, 91(6):811, 1940.

[104] B.-S. Xie. Recovery of Individual Head-Related Transfer Functions from a Small Set of Measurements. *The Journal of the Acoustical Society of America*, 132(1):282–294, 2012.

[105] C. Xu, T. Tan, Y. Wang, and L. Quan. Combining Local Features for Robust Nose Location in 3D Facial Data. *Pattern Recognition Letters*, 27(13):1487–1494, 2006.

[106] P. Yan and K. W. Bowyer. An Automatic 3D Ear Recognition System. In *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 326–333, Noth Carolina, USA, 2006. IEEE.

[107] P. Yan and K. W. Bowyer. Biometric Recognition Using 3D Ear Shape. *IEEE transactions on pattern analysis and machine intelligence*, 29(8):1297–308, 2007.

[108] S. Yang, S. A. Scherer, and A. Zell. Visual SLAM for Autonomous MAVs with Dual Cameras. In *IEEE International Conference on Robotics and Automation*, pages 5227–5232, 2014.

[109] S. Yang, S. A. Scherer, and A. Zell. Robust Onboard visual SLAM for autonomous MAVs. *Intelligent Autonomous Systems*, pages 361–373, 2016.

[110] M. Yokota. Head and Facial Anthropometry of Mixed-Race US Army Male Soldiers for Military Design and Sizing: A Pilot Study. *Applied Ergonomics*, 36(3):379–383, 2005.

[111] Z. Zhuang and B. Bradtmiller. Head-and-Face Anthropometric Survey of U.S. Respirator Users. *Journal of Occupational and Environmental Hygiene*, 2(11):567–576, 2005.

[112] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis. HRTF Personalization Using Anthropometric Measurements. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 157–160, New Paltz, USA, 2003. IEEE.

[113] S. Zuffi, A. Kanazawa, and M. J. Black. Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.