Motion in Depth

Complementary Sensing of Human Body Motion with IMU and Depth Data

DISSERTATION

zur Erlangung des Grades eines Doktors der Ingenieurwissenschaften

vorgelegt von

M. Sc. Jochen Kempfle

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät
der Universität Siegen

Siegen 2023

Betreuer und erster Gutachter

Prof. Dr. Kristof Van Laerhoven
Universität Siegen

Zweiter Gutachter

Prof. Dr. Jöran Beel
Universität Siegen

Tag der mündlichen Prüfung

22. April 2024

DECLARATION

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Freiburg, 2. Oktober 2023

Jochen Kempfle

ABSTRACT

This dissertation is about sensing and utilizing various kinds of human body motion, ranging from large-scale limb movements down to subtle respiratory motion of the upper body. One key idea hereby is to simultaneously capture different types of limb and body movements from two inherently different input modalities and to combine that data in a complementary way. The used modalities are a depth camera on the one hand and body-worn inertial measurement units on the other hand. It will be shown how such a complementary sensing approach can be achieved and how it can lead to the emergence of completely new applications that cannot easily be accomplished by the respective modalities on their own. For this, a novel method is proposed that allows the matching of the motion data as obtained from a wearable inertial sensing device and from the pose estimation on a depth camera's video stream. As a result, this method allows the identification of the person and limb the wearable device is worn on within the depth footage of an observing depth camera in the surroundings. Such an identification allows both modalities to establish a communication channel where person related data can be transmitted to the correct person and device within multi-person scenarios. An exemplary application would be indoor localization in places that use surveillance cameras anyway, for instance in an airport. Here, the position of multiple persons within the field of view of a camera in the surroundings could be tracked and, if a user decides to use this feature and enables it on a wearable device, its position can on demand be transmitted to the correct person and device.

To facilitate such applications, furthermore, a novel compression scheme for quaternion-based motion data will be elaborated. It has the purpose to reduce the amount of data to be transmitted in order to reduce energy consumption and to save precious bandwidth. The compression will be achieved by a novel piecewise linear approximation algorithm and relies on the fact that, similar to computer animations, only body postures at key positions, so-called keyframes, need to be stored or transmitted while the overall motion can be interpolated from these.

Finally, depth data will thoroughly be evaluated towards its usage for the remote sensing of respiration by measuring the subtle movements of the upper body caused by the elevation of the chest and abdomen during breathing. For this, a novel depth-based algorithm to robustly monitor human respiration from a distance is proposed. This method does not require any physical body contact, works reliably in distances up to 4 meters and, in contrast to available approaches, even works in the presence of occlusions and upper body movements as for instance are introduced while standing and keeping balance. This will be validated by comparing the proposed algorithm to a commercial respiration belt in a validation study. Furthermore, this method as well as the most common state-of-the-art depth-based respiration estimation methods will be compared on a thorough user study where a selection of the most relevant parameters that influence the respiration estimation are evaluated in depth.

ZUSAMMENFASSUNG

Ziel dieser Dissertation ist die Erfassung und Nutzung verschiedener Arten menschlicher Körperbewegungen, welche von ausladenden Bewegungen der Gliedmaßen bis hin zu subtilen Bewegungen des Oberkörpers während der Atmung reichen können. Eine Idee dieser Dissertation hierbei ist es die verschiedenen Körperbewegungen durch zwei inhärent unterschiedliche Eingabemodalitäten gleichzeitig zu erfassen und anschließend in einer sich gegenseitig ergänzenden Weise zu kombinieren. Die verwendeten Eingabemodalitäten sind zum einen eine Tiefenkamera und zum anderen am Körper getragene Inertialsensoren. Es wird gezeigt, wie solch eine sich gegenseitig ergänzende Kombination beider Modalitäten aussieht und wie dies zur Entstehung völlig neuer Anwendungen führen kann, welche durch die jeweiligen Modalitäten allein nicht ohne weiteres erreicht werden könnten. Zu diesem Zweck wird eine neuartige Methode vorgestellt, die es ermöglicht, die Datenströme beider Modalitäten abzugleichen. Als Resultat können sowohl die Person als auch das Körperteil, an dem der Inertialsensor getragen wird, innerhalb eines Tiefenbildes einer in der Umgebung angebrachten Tiefenkamera identifiziert werden. Dies ermöglicht es beiden Modalitäten einen privaten Kommunikationskanal aufzubauen, über den personenbezogene Daten an die richtige Person bzw. das richtige Endgerät selbst in Szenarien mit mehreren Personen übertragen werden können. Eine beispielhafte Anwendung wäre die Positionsbestimmung in Innenräumen. In Umgebungen an denen ohnehin Überwachungskameras zum Einsatz kommen, beispielsweise in einem Flughafen, könnten die Positionen verschiedener Personen im Sichtfeld einer Kamera erfasst und an das jeweils richtige Gerät übermittelt werden, vorausgesetzt der jeweilige Nutzer entscheidet sich dafür und aktiviert diese Funktion auf seinem tragbaren Endgerät.

Um derartige Anwendungen zu erleichtern wird weiterhin ein neuartiges Kompressionsverfahren vorgestellt, das für die Komprimierung von auf Quaternionen basierenden Bewegungsdaten konzipiert ist. Dieses dient zur Reduktion der zu übertragenden Datenmenge um damit sowohl den Energieverbrauch als auch die Bandbreitennutzung zu senken. Die Kompression wird durch einen neuartigen Piecewise Linear Approximation-Algorithmus erreicht und beruht auf dem Konzept, dass, ähnlich wie bei Computeranimationen, nur bestimmte Körperstellungen, so genannte Keyframes, gespeichert oder übertragen werden müssen, während die Gesamtbewegung des Körpers aus diesen interpoliert werden kann.

Schließlich wird die Verwendung einer Tiefenkamera eingehend zur Messung der menschlichen Atmung aus der Distanz untersucht. Das darunterliegende Prinzip hierbei beruht auf der Erfassung der subtilen Bewegungen des Oberkörpers, die durch das Anheben von Brust und Bauch während der Atmung verursacht werden. Zu diesem Zweck wird ein neuartiger Algorithmus vorgestellt, der eine robuste Messung der menschlichen Atmung unter Verwendung von Tiefendaten erzielen kann. Diese Methode erfordert keinen physischen Körperkontakt, funktioniert zuverlässig aus Entfernungen von bis zu 4 Metern und funktioniert im Gegensatz zu verfügbaren

Ansätzen auch bei Verdeckungen und leichten Bewegungen des Oberkörpers, wie sie beispielsweise beim Stehen und Halten des Gleichgewichts auftreten können. Dies wird überprüft indem die vorgestellte Methode in einer Benutzerstudie mit einem kommerziellen Atemgürtel verglichen wird. Darüber hinaus werden diese Methode sowie die gängigsten tiefenbasierten Methoden zur Erfassung der Atmung in einer umfassenden Benutzerstudie verglichen, in der eine Auswahl der wichtigsten Parameter, die die Messung der Atmung aus der Distanz beeinflussen können, eingehend evaluiert werden.

PUBLICATIONS

The following first author contributions have been made during the dissertation:

- Jochen Kempfle and Kristof Van Laerhoven. "Human posture capture and editing from heterogeneous modalities." In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 489–494. ISBN: 9781450353786. DOI: 10.1145/3152832.3157812
- Jochen Kempfle and Kristof Van Laerhoven. "PresentPostures: A Wrist and Body Capture Approach for Augmenting Presentations." In: *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, Mar. 2018. DOI: 10.1109/percomw.2018.8480155
- Jochen Kempfle and Kristof Van Laerhoven. "Respiration Rate Estimation with Depth Cameras: An Evaluation of Parameters." In: *Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction*. iWOAR '18. New York, NY, USA: Association for Computing Machinery, 2018. ISBN: 9781450364874. DOI: 10.1145/3266157.3266208
- Jochen Kempfle and Kristof Van Laerhoven. "Towards Breathing as a Sensing Modality in Depth-Based Activity Recognition." In: *Sensors* 20.14 (July 2020), p. 3884. ISSN: 1424-8220. DOI: 10.3390/s20143884
- Jochen Kempfle and Kristof Van Laerhoven. "Quaterni-On: Calibration-free Matching of Wearable IMU Data to Joint Estimates of Ambient Cameras." In: *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2021, pp. 611–616. DOI: 10/kt2k
- Jochen Kempfle and Kristof Van Laerhoven. "Breathing In-Depth: A Parametrization Study on RGB-D Respiration Extraction Methods." In: *Frontiers in Computer Science* 3 (2021). ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.757277

Additionally, the following co-author contributions have been made:

- Chaithanya Kumar Mummadi, Frederic Philips Peter Leo, Keshav Deep Verma, Shivaji Kasireddy, Philipp M Scholl, Jochen Kempfle, and Kristof Van Laerhoven. "Real-time and embedded detection of hand gestures with an IMU-based glove." In: *Informatics*. Vol. 5. 2. MDPI. 2018, p. 28. DOI: 10.3390/informatics5020028
- Florian Grützmacher, Jochen Kempfle, Kristof Van Laerhoven, and Christian Haubelt. "fastsw: Efficient piecewise linear approximation of quaternion-based orientation sensor signals for motion capturing with wearable imus." In: *Sensors* 21.15 (2021). ISSN: 1424-8220. DOI: 10.3390/s21155180
- Steffen Brinkmann, Jochen Kempfle, Kristof Van Laerhoven, and Jonas Pöhler. "Evaluation of a Depth Camera as e-Health Sensor for Contactless Respiration Monitoring." In: *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2023, pp. 136–141. DOI: 10.1109/PerComWorkshops56833.2023.10150271

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisor Prof. Dr. Kristof Van Laerhoven for always having an open ear, for his unbreakable motivation, open-minded attitude and great guidance, and for his great ideas that helped shape this dissertation. I also thank Prof. Dr. Jöran Beel for taking the time and being the co-examiner of this thesis. Thanks also go to Alex, Flo, Jonas, and Marius from the UbiComp group for many good conversations and discussions that often led to great ideas, as well as to all other PhD students of the university of Siegen I spent some good time with. Dear participants of my user studies, thank you for your time, this thesis would not have been possible without you! A big thank you also goes to Christoph for always having the time and passion to set up any kind of technical request, be it a database for students, a server instance for a project group, or simply a workstation in the lab. Also, I would like to thank all my friends that supported me during the thesis and that were always within reach despite the great distance of up to more than 400 km to Siegen. The same applies to my family and especially my nieces Ida and Ella with which I spent far to few time during this thesis. Special thanks go to Phil and Juliane that both agreed to carefully read through this dissertation and helped improve it. Last but not least, another really big thank you goes to my girlfriend Juliane for her patience and support over all the time. It would have been different and certainly less enjoyable without you.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AAD	Average Angular Deviation
AHRS	Attitude Heading Reference System
BLE	Bluetooth Low Energy
bpm	breaths per minute
CPLR	Connected Piecewise Linear Regression
EKF	Extended Kalman Filter
fastSW	Fast Sliding Window
FFT	Fast Fourier Transform
GCC	GNU Compiler Collection
ICR	Inverse Compression Ratio
IMU	Inertial Measurement Unit
IQR	Interquartile Range
MoCap	Motion Capturing
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PLA	Piecewise Linear Approximation
ROI	Region of Interest
SF	Swing Filter
SLERP	Spherical Linear Interpolation
SNR	Signal-to-Noise Ratio
SSR	Sum of Squared Residuals
SVM	Support Vector Machine
SW	Sliding Window
ToF	Time of Flight
UDP	User Datagram Protocol
UKF	Unscented Kalman Filter
WCET	Worst-Case Execution Time

1 INTRODUCTION

1.1 MOTIVATION

Sensing human body motion plays a central role in a variety of different applications. Most prominently, it is represented in the modern film and gaming industry, where the most advanced Motion Capturing (MoCap) systems are used to not only simplify and accelerate the process of character animation, but also to create vivid, lifelike animations. In the field of human-computer interaction, the line between detecting user input and sensing motion increasingly vanishes. One can observe a trend from simply pressing a button over detecting a touch or a touch gesture, towards remotely detecting an actual gesture, for instance to take a selfie with a smartphone's camera by giving a hand sign, and applications like virtual reality would not be possible without capturing human body motion.

Apart from such well known applications, sensing human motion plays an even more important role in many fields of research and science, but also in applications of daily living. Many user studies acquire and digitize human body movements to examine, assess, compare, or parameterize the motion of the full body or specific limbs [35, 48]. Examples range from collaborative robotics [39, 162] as well as the assessment of the safety or efficiency of motion sequences at work [20, 42, 175], over applications such as sports and fitness [86], up to health sciences [67, 133], to name a few. Consequently, monitoring body movements in the medical sector e.g. for gait analysis [31, 68, 142] or during sleep assessment [37] can provide better insights and aid in making a diagnosis. In daily applications, step detection or fall detection is deployed in many smart devices, where such wearable motion sensing technology can for instance also be used to increase the physical activity of older adults [33]. Sensing the motion of the hands and fingers enables gesture detection [120] and in combination with the detection of posture, gaze, and facial expression, even simplified sign language recognition can be achieved [32]. Even in automotive applications, the motion of pedestrians is assessed to predict their behavior [141]. While some fields can benefit from sensing limb and body motion but do not necessarily require it, other fields do heavily depend on capturing motion. One example hereby is the field of human activity recognition, where all kinds of limb and body motions are being detected and characterized. Examples range from detecting household activities over different types of locomotion and transportation up to the recognition of sports and leisure activities.

Depending on the task and the application, be it activity recognition or something else, either the body as a whole has to be monitored, or only specific movements or specific limbs need to be observed. Accordingly, there exist numerous techniques and devices to sense human body motion. These can be divided into two major groups: On-body sensing on the one hand and remote sensing on the other hand. In short, on-body sensing typically is achieved by body-worn inertial sensors, so-called

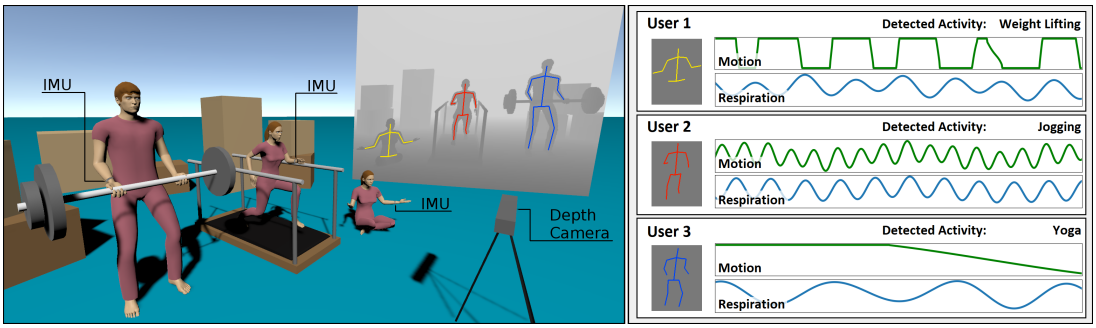

Figure 1.1: Example application where users wear a fitness tracker or a smartwatch with an integrated IMU and are observed by a depth camera while doing leisure activities such as sports or meditation exercises (left). The data streams from both modalities can be combined in a complementary way when the limb and person a wearable device is worn on can be identified and matched to the correct person and limb as observed from the depth camera. After such a matching, it is possible to asses a variety of different body parameters, which then can be transferred back to the wearable device of the respective person. Such data can include full body posture, limb movements, or user position, for instance to enable activity recognition or indoor localization, but also physiological body signals, such as respiration, as can be obtained from depth data by measuring the tiny changes of chest elevation during breathing (right).

Inertial Measurement Units (IMUs), and remote sensing typically is achieved using some sort of camera or camera system in the surroundings.

In the field of activity recognition and applications thereof, motion data preferably is being captured from body-worn inertial sensors due to their high mobility and low cost. The most important reason, however, is that these sensors are built-in into modern smartphones, smartwatches, or similar wearable devices. These devices have the advantage of being accepted and used by the broad mass of the population on a daily basis, so there is no need to deploy specialized hardware to people interested in using activity recognition because they can just use their wearable smart device to benefit from it. This, on the other hand, comes with a drawback: People do only have a limited amount of wearable devices that furthermore are only worn at a few specific body positions, depending on the device and the personal preferences of the user. Motion data thus can only be captured from one or a few specific body parts, for instance the left or right forearm in the case of a smartwatch. This effectively limits the set of possible fields of applications, e.g. for activity recognition, from the outset. Attaching multiple sensors to a body is simply not practical for everyday use.

Optical motion capturing, on the other hand, typically allows to track the full body, even of multiple persons. The use of an expensive MoCap system comprising highly optimized cameras as well as the use of fiducial markers hereby is not necessarily required. Advances in machine learning led to systems that can achieve multi-person pose estimation in real-time using only a single standard RGB-camera. An alternative to such single RGB-camera systems that has long been proven effective is the use of a single depth camera. Depth data has the convenience that three-dimensional data can relatively easy be retrieved from the observed scene, which led to its successful application for highly performant and accurate 3D pose estimation early on. A prominent

example hereby is the Kinect. It is a consumer grade depth camera that is able to efficiently achieve three dimensional pose estimation in real-time and at low cost. For this reason, it can be found in many scientific studies as well as in many other applications that rely on 3D pose estimation. Since the introduction of the Kinect, depth cameras have become smaller, cheaper, and more precise, and nowadays are not only deployed as stand-alone devices, but with increasing regularity even are integrated into many consumer grade devices, such as smartphones. This makes depth cameras not only a viable alternative for traditional optical motion capturing techniques, but potentially also opens up a range of new applications: Accurate estimation and tracking of human body motion can theoretically be achieved by any user of such a device on a daily basis. The question then arises, how this availability of depth sensing technology could be used and how such applications could look like. This is where this work jumps in. This dissertation has the aim to provide an in-depth analysis of the different types of human body motion present in depth data, ranging from large-scale limb movements down to subtle movements of the upper body during respiration. It shows how to exploit this data for complementary motion sensing or to monitor a person's respiration from a distance and without the requirement of any physical body contact.

Complementary motion sensing in this case means the simultaneous observation of a human's limb movements from one or a few body-worn inertial sensing devices and a depth camera, with the aim to fuse both observations in order to achieve an optimized end result. An obvious use-case would be the optimization of MoCap systems, where the depth-based pose estimation of the full body for instance could be used to complement sparse motion data from body-worn sensing devices, or vice-versa, optical pose data could be complemented by motion data from body-worn devices, e.g. in case of occlusion or when the observed person is out of frame. This easily extends to the field of activity recognition, where a wearable device can benefit from becoming aware of the user's full body pose and motion. The range of possible applications thereby is not restricted to such use-cases, but, as will be seen later on, complementary motion sensing can eventually lead to the emergence of completely new applications that even extend into the domain of daily living. This, however, requires that the sensing devices of both modalities are able to communicate with each other on the fly and without knowing each other beforehand. Thus, a core requirement would be that both sensing devices or sub-systems are able to identify each other. To date, it remains widely unclear how such an identification could look like and how complementary motion sensing could be established in general.

Capturing respiration from a distance and without any body contact, on the other hand, can be of advantage in many different applications. In the medical sector, a patient could automatically be observed by such a system, either for non-obtrusive long-term monitoring both at home or in hospital, or to safe precious time of doctors and nurses that for some patients and diseases need to manually assess respiratory rate on several minute-long sessions. Also, not requiring physical body contact can arguably be beneficial for both, the patient and the doctor or nurse in charge. Since breathing can either not or only indirectly be retrieved on smartwatches or similar body-worn devices, i.e. by calculating heart rate variability on the signal obtained from photoplethysmography, the remotely measured respiration signal could also

be sent to the respective end device, which then can benefit from being aware of an additional physiological signal. In combination with the non-obtrusive character of remote respiration estimation, this makes remote respiration estimation a viable alternative for a range of other applications as well. In the fields of fitness and sports, an athlete can assess its respiration for instance as a performance measure, and in meditation, where breathing exercises are very common, the respiration can be observed in order to provide instant feedback, either by a supervisor or by the system itself. As a last example, and as will be seen later in this work, it furthermore can serve as a complementary modality for activity recognition or be used in medical applications such as e-health and telemedicine.

The principle of depth-based respiration estimation hereby is based on measuring the subtle movements of the upper body caused by inhalation and exhalation during a respiration cycle. These movements, although being tiny in scale, can become visible on the depth data of a human's upper body when being observed by a depth camera. For a reliable measurement, however, the respiration detection has to be robust against any disturbing influences, including measurement-related influences such as noise as well as disturbing influences caused by the person under observation. Person-related influences range from different clothing styles over user-specific breathing patterns, e.g abdominal or thoracic breathing, up to conscious or unconscious movements of the upper body, e.g. while standing and keeping balance or while moving the arms and partially occluding the upper body, but also other factors can play a role. These influences are neglected by any depth-based respiration estimation methods proposed so far and user studies remain far from realistic: Participants are required to either lie down or sit still. Furthermore, available literature lacks a comparison of available depth-based respiration estimation techniques as well as a systematic study of relevant parameters that influence the measurement of the respiration signal.

If IMU and depth data already are used as complementary modalities, for instance to assess limb and body movements, such a remote respiration estimation can easily be integrated into the complementary sensing approach without the need for any further equipment. Figure 1.1 depicts an exemplary use-case of such a complementary use of both modalities in a sports and meditation scenario.

1.2 CONTRIBUTIONS

One key idea of this dissertation is to use a depth camera to capture the entire body and its movements in certain situations, to identify the respective device worn by the respective person and to augment the externally captured data to this device. The body-worn device then can complement its knowledge of the limb it is worn on with externally captured body data, such as posture and limb movements. Moreover, the external device can as well provide any kind of meta-data, including externally captured physiological data such as respiration, directly observable data such as a user's location, or more general kinds of data such as the context of a situation. Overall, this opens up completely new possibilities in a variety of different areas, ranging from indoor localization to activity recognition. The key element for such applications is an algorithm that can be used to identify which person in a depth video recording

is wearing which end device on which part of the body. This algorithm is based on matching the motion detected by the respective inertial sensor with the body movements of all visible persons captured in the video. One goal of this dissertation thus is the development of such an algorithm and its evaluation in a realistic scenario. It will be embedded in a more general approach on how to work with multi-modal sensor data as obtained from an external depth camera and body-worn IMUs in order to achieve complementary motion sensing.

Since any application that works with complementary motion sensing would require the transmission of a considerable amount of motion data, quaternion-based motion data will thoroughly be analyzed towards a viable compression scheme, eventually leading to a novel and efficient compression method. It relies on the circumstance that human body motion typically does not involve the movement of all body parts at the same time, nor that they are moved in completely random directions at completely random paces. To obtain the overall motion of a human body it is sufficient to store or transmit only the body postures at specific key positions. Similar to the field of computer animation, where so-called keyframes are used, the overall motion then can be interpolated from these. The proposed compression scheme is based on a Piecewise Linear Approximation (PLA) technique that is specifically tailored to work in environments with limited computational resources such that quaternion-based motion data can directly be compressed on the sensor system.

Another aspect of this dissertation will be the analysis of the depth data for the contact-less and distant detection of the breathing of a person in the field of view of a depth camera. Depth-based respiration estimation methods per se are not novel, but, as discussed in the previous section, most of them remain far from realistic and particularly neglect person-related influences. To overcome these limitations, a novel method for robust depth-based respiration estimation is proposed that does not require users to lie down or to sit still and that can handle situations where the observed person is standing freely while regularly occluding its upper body with an object held in the hand. As already mentioned above, available literature lacks a comparison of current state-of-the-art depth-based respiration estimation techniques and the influence of relevant parameters on the measurement of the respiration signal is widely unknown. So, in addition to the implementation of such a robust algorithm, a systematic evaluation of parameters relevant for depth-based respiration estimation is carried out and their influence on the obtained respiration signal is determined on the basis of objective performance measures. This will be done for the presented algorithm as well as for the currently existing methods. Among the parameters to be investigated are the optimal area for extracting the respiration signal, the distance to the camera, the breathing rate, and whether the user is sitting, standing, or even occluding its upper body, but also other user-specific characteristics, such as the gender or the individual clothing style will be considered.

The most important contributions of this dissertation can be summarized as follows:

- A study on the tracking performance of the wrist joint in a complementary motion sensing scenario where optical and inertial MoCap data is used in combination, accompanied with a discussion of requirements and considerations for the handling of such a multi-modal sensor input.

- A novel method to identify the person and limb a wearable inertial sensing device is worn on within such complementary motion data by comparing and matching limb movements from both modalities.
- An evaluation and discussion of the key requirements for applying PLA algorithms to quaternion-based motion data, resulting in a novel PLA algorithm for the efficient approximation of quaternion-based orientation sensor signals in environments with limited memory and computational resources.
- A novel, depth-based method for the remote and contact-less monitoring of human respiration that is robust against several challenging conditions including user motion and partial occlusions of the user's upper body.
- Two publicly available respiration datasets for a validation and a systematic parameter evaluation of depth-based respiration estimation methods with a combined length of more than 11 hours of respiration data, comprising 422 unique recordings from 24 different participants that were taken from different distances, respiratory rates, and activities.
- An in-depth evaluation of to-date unknown influences of important key parameters on the most common state-of-the-art depth-based respiration estimation methods as well as on the proposed approach. Examined parameters are the observed torso region, whether the user is sitting, standing, or standing with regular self-occlusions, the distance to the depth camera, the respiratory rate, the gender, and user-specific influences.
- An experiment-driven exploration of possible applications for depth-based respiration estimation.

1.3 OUTLINE

This dissertation is structured into four main parts. In Section 2 (*Related Work*), the literature related to complementary motion sensing, piecewise linear approximation, and remote respiration estimation is described and the work shown here is situated amongst the current state of the art in these fields. Section 3 (*Complementary Motion Sensing*) provides a discussion and a case study on how to handle multi-modal sensor input for capturing human body motion using inertial and optical motion data. After that, a novel technique is introduced that allows to identify the person and limb a wearable inertial sensing device is worn on within a depth video by matching inertial to depth-based limb motion estimates. Section 4 (*Compressing Motion Data with Piecewise Linear Approximation*) discusses and evaluates the key requirements for applying PLA algorithms to quaternion-based motion data and shows how an efficient approximation of such data can be achieved using a novel PLA algorithm specifically tailored to environments with limited memory and computational resources, as can be found in many sensor devices used for capturing motion. Finally, Section 5 (*Remote Respiration Estimation*) introduces a novel method that is able to remotely monitor human respiration from depth data. The proposed method is robust against user motion and partial occlusions of the upper body as will be shown

in a validation study. Furthermore, an in-depth evaluation of to-date unknown influences of important key parameters on the most common depth-based respiration estimation methods is conducted using an extensive respiration dataset that specifically was recorded for both these studies. It is followed by an exploration of possible applications of remote respiration estimation in the fields of activity recognition and e-health. A summary and conclusion can be found in Section 6.

2 RELATED WORK AND STATE OF THE ART

2.1 COMPLEMENTARY MOTION SENSING

Motion capturing is successfully used in a variety of applications, with the film and games industry being a prominent example. Here, optical and IMU-based full body tracking systems already are deployed industrially [1, 2, 143]. With the advent of depth cameras, such as the Kinect, and by utilizing the depth information to first estimate body parts and afterwards compose the posture of the observed subjects [150], also consumer-grade optical Motion Capturing (MoCap) systems entered the market. For the Kinect v2, the accuracy of the joint estimation can be comparable to a standard optical motion capture system, given a controlled body posture, for instance standing upright and exercising arms while facing the depth camera [126]. More recently, advances in machine learning led to methods that allow human pose estimation in real-time from single RGB images [25, 102, 116, 117].

For IMU-based systems on the other hand, recent advances enabled full body human pose estimation from sparse inertial measurements using only 6 Inertial Measurement Units (IMUs) attached to the wrists, lower legs, back and head [64, 164] or with even less body-worn IMUs in phones, watches, or earbuds [119], but at the cost of lower accuracy. In principle, it would also be possible to attach a miniature depth camera on a limb and to use the depth information to track and locate the moving sensor by simultaneously mapping the environment [123] and thus to infer the limb's motion. For full body pose estimation this approach, however, is not feasible due to cost, size, weight, and energy consumption, especially when compared to IMUs. Furthermore, there also exist methods that do not directly aim for pose estimation, but instead focus on extracting virtual on-body sensor data from RGB video footage in order to simulate body-worn sensors, for instance for human activity recognition purposes [91].

Optical and inertial motion capture systems are known to have their specific strengths and weaknesses. Self-occlusion by the person under observation and occlusion by nearby structures, as well as adverse lighting conditions tend to hamper an accurate body posture recognition for optical MoCap systems [135]. Also, full-fledged optical systems tend to be less flexible to be moved at different locations, and their setup effort and costs tends to be higher than wearable inertial measurement solutions. IMU-based systems on the other hand are not as accurate, cannot be used to track a user's position, and suffer from noise and sensor drift, but are not restricted to certain working volumes. Integrating the IMU sensor data into a biomechanical model and modelling the sensor to segment offset as demonstrated in [85] or [172] does increase the overall accuracy [143]. Soft tissue deformation and a loose or erroneous sensor attachment still poses a problem for IMU-based as well as for optical, marker-based MoCap systems [36]. Carefully placing the sensors or fiducial markers and accurately accessing the various calibration parameters therefore is a

vital requirement to minimize these effects. Also, a well performed camera and IMU calibration plays an important role to achieve a good measurement accuracy in the first place, as demonstrated in [139].

Looking at the strengths and weaknesses of both sensor modalities, it can be seen that both complement each other well and in recent years, some examples have shown how the weaknesses in one modality can be addressed by another. In [30] for instance, their introduced fusion approaches led to improvements of 2% to 23% of the recognition rate in action recognition when using features from depth images and from the accelerometer signal of a body-worn inertial sensor in combination, as compared to using each sensor individually. IMU data has also successfully been fused with non-optical modalities, for instance with an electric potential sensor to improve scratch detection in health applications [73]. In this work, however, the focus lies on combining inertial and optical motion data and such applications are left out. Furthermore, there exist particular forms of fusing inertial with optical motion data, such as placing a RGB camera next to an IMU on the same board and stabilizing the IMU predictions by estimating the motion from the movement of the surroundings as seen from the camera [47]. Such methods are also excluded. The complementary sensing approach envisioned in this work focuses on observing the movements of an user from an external depth camera and combining that data with motion data obtained from one or more IMUs worn by the same user.

In the following, a compilation of various methods to combine inertial and optical MoCap modalities are presented, separated into RGB or depth-based methods. After that, a short overview is given on how to locate IMUs in video footage based on their measurement data. For further reading, recent surveys about fusing inertial with optical data for human pose estimation can be found in [96] and [124], and a survey about fusing both modalities for human action recognition can be found in [106].

2.1.1 *Combining Inertial and Optical Motion Data*

2.1.1.1 *Single Camera RGB-Based Methods*

Some proposed systems combine inertial data with video footage of a single RGB camera, for instance to stabilize the position ambiguity of inertial pose estimation by detecting the user's foot position on the ground [74] or, the other way round, to use inertial data to succeed in difficult situations, where RGB-based pose trackers fail due to insufficient data, such as non-frontal poses or occlusions, for instance by combining a generative and a discriminative tracker to retrieve closest poses from a database [61]. Some works also use fiducial markers that can be tracked by a single RGB camera. These are either placed on the inertial sensor modules to compute the drift-free orientation of the modules through visual-inertial fusion [95], or are placed on the body and, after determining their exact location on the body segments during a calibration phase, all measured inertial and optical quantities are fused into a bio-mechanical model using a constrained extended Kalman filter [108]. Above methods all require a fixed camera, but it also is possible to estimate 3D human pose accurately on a person equipped with body-worn IMUs and filmed by a moving RGB camera in the wild [163].

2.1.1.2 *Multi Camera RGB-Based Methods*

Instead of relying on only a single RGB camera, also methods that use multiple RGB cameras in combination with IMU sensors are proposed. For instance, by matching the 3D surface mesh of an actor, as recorded prior to the experiment, to 2D image contours coming from a set of calibrated and synchronized RGB cameras in the surroundings, it is possible to estimate the pose of the actor, which then can be improved upon by minimizing the geodesic or the chordal distance between the estimated limb orientations and the orientation data from body and limb mounted inertial sensor units [113]. On a multimodal dataset [112] specifically recorded for that study, it was shown that a hybrid approach that combines information of optical and inertial modalities can significantly improve the tracking quality by resolving inherent ambiguities when reconstructing a 3D pose from 2D video data, for instance errors arising from rotationally symmetric limbs and noisy visual cues [113].

Another method that does not need a surface mesh or full body model to estimate 3D human pose by fusing multi-viewpoint video with IMU sensor data is presented in [159] and [50]. In both works, a 3D convolutional neural network is trained to learn a pose embedding from the video data and a LSTM model is incorporated within the pose stream and the forward kinematic solve of the IMU data before both are fused in a fully connected layer. Both complementary data sources reportedly allow for ambiguities to be resolved, leading to an improved accuracy.

A real-time full-body motion capture system based on a sparse set of IMUs and images from two or more RGB cameras is presented in [110]. It requires no optical markers, but incorporates constraints from the IMUs, cameras and a prior pose model into the proposed optimization-based framework and allows the full 6-DOF motion to be recovered, including axial rotation of limbs and drift-free global position as tested in indoor and outdoor scenes. Sorting and assigning OpenPose's [25] 2D keypoint detections into corresponding subjects furthermore facilitates multi-person tracking and rejection of any bystanders in the scene [109].

In [177], a so-called Orientation Regularized Network is used to improve 2D pose estimation accuracy from multi-view images and a few IMUs attached at a person's limbs. The multi-view 2D poses then are lifted to the 3D space by minimizing the projection error from 2D to 3D by using a so-called Orientation Regularized Pictorial Structure Model.

2.1.1.3 *Depth-Based Methods*

Some works show that IMU-based systems can be cost-effective and dynamically deployable, yet face calibration and floating artifacts for hip-joint rooted methods [174, 178]. Indoor magnetic disturbances are also known to affect the IMU-based units' accuracy, leading to a variety of research efforts to characterize and compensate for this, as for instance summarized in a survey and collection of methods [97]. A RGB-D sensor, such as the Kinect, on the other hand, not only has to deal with occlusion, but also with some specific weaknesses such as being unable to correctly track a human from the side or from the back (see [131] and [128] for surveys on the Kinect abilities compared to a gold standard Vicon 3D motion capture system).

Destelle et al. [38] have shown how combining an RGB-D sensor with wearable inertial measurement units improves motion capture by tracking the initial calibration pose and subsequently the body's position with the Kinect and the limb movements with the IMUs. Utilizing RGB-D sensors for estimating the sensor to segment offset leads to superior motion capture results compared to estimating them by hand, relying on the correct sensor placement, or executing specific calibration movements [28]. Furthermore, determining each IMU's sensor drift using standard system identification methods with the respective Kinect's joint data output increases the overall long-time accuracy, even when the captured body no longer is tracked by the Kinect [24]. By using a Kalman filter to fuse the Kinect's joint data with IMU orientation data as calculated via the Madgwick filter [105], it is possible to track the movement of limb joints precisely and almost drift-free, since using the Kinect's absolute position information can compensate for the drift [75]. In [158], an Unscented Kalman Filter (UKF) based fusion of IMU and Kinect joint data to compute robust hand position information is compared to the double integration of the IMU's internal sensors and to IMU internal sensor fusion with geometrical constraints. Experimental results show that in contrast to the two approaches solely based on the IMU data, the proposed IMU and Kinect fusion method can provide drift-free and smooth results and that it is able to achieve better results than using the Kinect alone. An UKF based fusion approach also was developed in [11]. It fuses the orientation data of the Kinect and body-worn inertial sensors for human arm motion tracking, specifically to compensate for the drift of inertial sensors and the occlusion of joints as seen from the Kinect. Compared to only using either Kinect or inertial sensor data, the errors could be reduced by almost 50%. Also [69] and [38] show how IMU-based tracking can improve the Kinect data, especially under occlusion. Using a personalized articulated human mesh model computed from a single depth image and two wrist-worn IMUs to provide additional clues for the arm tracking of the Kinect v2, e.g. during an occlusion, can reduce upper-limb joint position errors by 20% as compared to the Kinect's skeleton tracking alone [70]. With a sensor glove comprising multiple IMUs and passive visual markers, and a head-mounted stereo camera, also the estimation of hand and finger motion can be achieved by using a visual-inertial fusion algorithm that takes into account the hand anatomical constraints [93]. Another work proposes a hybrid motion tracking algorithm that only uses a single depth camera and sparse IMUs and that enables non-rigid surface reconstruction even for fast motions and challenging poses with severe occlusions, including the inner human body shapes of a clothed subject [179].

2.1.2 *Locating IMUs on Human Bodies*

Since the introduction of inertial sensors, several studies have been published that use inertial data to estimate where on the body these are likely attached. Early approaches to detect sensor placement from acceleration and gyroscope data during different activities have shown 100% accuracy for a walking activity, and up to 82% for several real-life activities [87, 89]. In [88], the authors further explored how sensor placement variations (head, wrist, torso, left breast pocket, and front and back trouser pockets) can influence human action recognition and how on-body placement of the

sensors can be detected. With their method, it typically takes some time (up to a few minutes) to reach the peak accuracy in detecting the correct sensor placement. A variety of classifier-based approaches have been proposed to see where inertial sensors are attached to: [7] uses a Support Vector Machine (SVM) to identify the location of 10 accelerometers on various parts of the body with an accuracy of up to 89%. Converging times, however, are not given and the method can not distinguish left from right body locations. A more frequent use of the right arm is assumed. In [169], a decision tree is trained on 17 inertial sensors placed on different limbs and achieves an accuracy of 97.5% in estimating the sensor placement. The setup requires a known sensor configuration and a walking pattern with sufficient arm movement (one participant with insufficient arm movement was excluded). Without knowing the sensor configuration, the accuracy drops to 75.9%. In [180], the sensor alignment and assignment on the lower body is estimated using deep learning. An accuracy of 98.57% is reported on the assignment classification using synthetic and real acceleration and gyroscope data for training. In [111], walking and non-walking accelerometer data from 33 participants, each wearing 5 accelerometers at ankle, thigh, hip, arm, and wrist, was recorded and the placement of each sensor estimated. Estimation was done through splitting the data in non-overlapping 10s windows and finding a walking motion with a SVM. If walking was detected, the location of the sensors is classified in a second step. Overall, a classification accuracy of up to 96.3% is reported using a majority voting strategy.

2.1.3 *Locating IMUs in Video Footage*

Only few works exist to date to identify an inertial sensor through its measurement data within simultaneously recorded video footage. By comparing an acceleration estimate of feature points in a RGB or RGB-D stream to the acceleration readings of an accelerometer that is attached to a limb or an object, it was possible to identify the accelerometer's location in the image domain [107, 154]. Also a person wearing an accelerometer could be identified in a video out of 3 walking people, and an accelerometer held in a moving hand could be identified out of 3 moving hands, however only on separate videos with only one person per video and only one of them wearing or holding an accelerometer, respectively [148]. IMUs have also been used to track multiple persons in a video by using a neural network that correlates the IMU's orientation and acceleration data with the movement of all currently visible persons in the video frame to identify which IMU belongs to which person in order to track that person even under heavy occlusions and if it is out of frame [62]. Also other works have used inertial data to improve tracking people in video data [71, 72], but the IMU has to be assigned manually to the person wearing it and the IMU data is only used when vision based tracking fails instead of using both modalities simultaneously.

2.2 COMPRESSING MOTION DATA WITH PIECEWISE LINEAR APPROXIMATION

Several Piecewise Linear Approximation (PLA) algorithms (or *segmentation algorithms*) have been introduced over the past decades, with the aim to reduce the amount of

data that has to be stored, transmitted, or further processed while keeping general trajectory information of the compressed signal. When directly deployed on a sensing device, such a PLA method can lead to an efficient operation of this device by effectively reducing memory and bandwidth requirements. Indirectly, also energy consumption of for instance sensor nodes that transmit their data wirelessly can be reduced this way [45, 55], given that the added workload for computing the PLA does not outweigh the energy saved on transmission or storage of the compressed data [53]. Both are desired effects, but for a deployment on a wearable sensor node, the used PLA method also has to meet certain requirements: Memory and computational resources usually are limited on such embedded platforms and the used method should be capable of approximating the signal online, i.e. as soon as a new sample arrives. Since PLA algorithms have never been used for the approximation of quaternion-based orientation sensor signals, in the following an analysis and comparison of existing PLA algorithms is performed. The focus hereby lies on their applicability on compressing quaternion-based motion data directly on the sensing device, i.e. in an environment with limited memory and computational resources. For this, a list of important factors that constrain such a use-case is compiled and used for the comparison (also see Table 2.1).

Two well-known PLA techniques are the Sliding Window (SW) and the Bottom Up (BU) algorithms [83]. Both were combined into the Sliding Window and Bottom Up (SWAB) algorithm by Keogh et al. [83], and with mSWAB [161] and emSWAB [16], further improvements to SWAB have been introduced. The latter, however, has an execution time that is magnitudes higher than other PLA algorithms.

PLA methods that can be executed in constant time and with constant memory consumption per processed sample and thus are able to run on architectures with limited resources are Swing Filter (SF) [43] and Connected Piecewise Linear Regression (CPLR) [52]. Both PLA algorithms determine the best fitting slope of PLA segments in a similar way, but differ in their error metric to decide on the termination condition of a segment. Furthermore, CPLR and SF extrapolate segment points from a regression line, which yields segment points that generally do not represent samples of the original signal, except by chance.

Another fast PLA algorithm was introduced by Lemire et al. in [94]. This algorithm, however, comes at a significant increase in memory consumption, which makes it unsuitable for a deployment on embedded systems or platforms with limited resources. Other PLA algorithms such as PLAMLiS [99] and its optimized variant [132] have at least a quadratic computational complexity to process a series of m samples and the complexity of processing a single sample, which is decisive for an online approximation, is not detailed for both.

SwiftSeg [49] is a segmentation framework based on polynomial least-squares approximation that can produce PLA segments as well, when first order polynomials are used, but it produces disconnected segments due to an intercept term in the linear regression. Another method that produces a mixture of connected and disconnected segments was introduced by Luo et al. [103]. It has a constant update time, but it is based on a buffer which limits the segment length and thus the data compression ability on memory constrained systems.

A comparison of aforementioned state-of-the-art PLA algorithms, including the proposed fastSW, is summarized in Table 2.1. The first column specifies the PLA methods and their respective origin, and the second column (OL) denotes their ability to process a sample online, i.e. at runtime. This capability is a necessary feature for motion capturing with wearable sensors, where the data is approximated on the microcontroller of the sensing device itself. The third column (CS) specifies if the PLA algorithm produces connected segments, a necessary feature for a seamless interpolation of the approximated signals on the receiver side. The fourth column lists if the PLA segment points are a subset of the original sensor samples or, in short, the preservation of sensor samples (POS). This is a crucial point for quaternion-based signals and will in detail be explained in Section 4. The fifth column (BB) denotes if the respective method is based on a buffer. A buffer does in most cases constrain the length of the segments, which lowers the achievable compression ratio. The sixth column comprises the error metric (EM) that is used for the error bound of the approximation as specified by the user. The time complexity (TCn) and memory complexity (MCn) of processing a single sensor sample with respect to the segment length n are contained in the seventh and eighth column, respectively. Ideally, both complexities are constant, i.e. in $O(1)$, because otherwise limited computational resources might constrain the maximum achievable compression ratio. Finally, the ninth and tenth column list the time complexity (TCm) and memory complexity (MCm) of processing the entire sequence of samples with respect to its length m. If applicable, the length of the buffer is assumed to be of length m, as this would yield the maximum compression ratio.

Table 2.1: Comparison of state-of-the-art PLA algorithms [54]. **OL**: Online applicability, **CS**: Produces connected segments, **POS**: Preserves original samples, **BB**: Requires a buffer, **EM**: Error metric, **TCn** and **MCn**: Time and memory complexity for processing a single sample with respect to the segment length n, **TCm** and **MCm**: Time and memory complexity for processing an entire sequence with respect to its length m, **SSR**: Sum of squared residuals error of a segment, **SAD**: Sum of absolute deviations of a segment, and ε: Absolute residual error per sample.

Algorithm	**OL**	**CS**	**POS**	**BB**	**EM**	**TCn**	**MCn**	**TCm** [1]	**MCm** [1]
BU [83] [2]	no	yes	yes	yes	SSR	$O(n^2)$	$O(n)$	$O(m^2)$	$O(m)$
SWAB [83]	yes	yes	yes	yes	SSR	$O(n^2)$	$O(n)$	$O(m^2)$	$O(m)$
mSWAB [161]	yes	yes	yes	yes	SSR	$O(n^2)$	$O(n)$	$O(m^2)$	$O(m)$
emSWAB [16]	yes	yes	yes	yes	SAD	$O(n^2)$	$O(n)$	$O(m^2)$	$O(m)$
PLAMLiS [99]	n/a	yes	yes	yes	ε	n/a	$O(n)$	$O(m^2 \log m)$	$O(m)$
PLAMLiS extension [132]	n/a	yes	yes	yes	ε	n/a	$O(n)$	$O(m^2)$	$O(m)$
SW [83] [2]	yes	yes	yes	yes	SSR	$O(n)$	$O(n)$	$O(m^2)$	$O(m)$
By Luo et al. [103]	yes	mixed	no	yes	ε	$O(1)$	$O(n)$	$O(m)$	$O(m)$
By Lemire [94]	(yes)[3]	n/a	no	yes	SSR	$O(1)$[4]	$O(n)$	$O(m)$	$O(m)$
SwiftSeg [49]	yes	no	no	(yes)[5]	SSR, ε, ...	$O(1)$	$O(1)$	$O(m)$	$O(1)$
CPLR [52]	yes	yes	no	no	SSR	$O(1)$	$O(1)$	$O(m)$	$O(1)$
SF [43]	yes	yes	no	no	ε	$O(1)$	$O(1)$	$O(m)$	$O(1)$
fastSW (proposed method, Sec. 4.3)	yes	yes	yes	no	SSR	$O(1)$	$O(1)$	$O(m)$	$O(1)$

[1] To provide worst-case bounds, the buffer lengths or the maximum segment lengths are assumed to be as large as the dataset itself, respectively.

[2] The original source does not get obvious from the literature.

[3] Although not explicitly stated in [94], the PLA algorithm of Lemire could be used for online processing.

[4] Although the calculation of line fit and error happens in an $O(1)$ step, it is based on a precalculated array of range sums for each sample of the sequence.

[5] In general, SwiftSeg is based on a buffer, but for the first order variant with segmentation and slope information, a buffer might not be necessary.

2.3 REMOTE RESPIRATION ESTIMATION

Systems that are able to monitor a user's breathing have been presented in the past for several applications and scenarios, with many health care and fitness-related aspects as a main motivation. To date, several approaches exist to measure respiration from a distance, either optically or with the use of RF-antennas. Optical methods hereby initially used standard RGB and near-infrared cameras and, more recently, increasingly take advantage of depth cameras as sensing devices. While RF-based and RGB- or infrared-based approaches for remote respiration estimation are an interesting research field on their own, in the following the focus mostly is set on depth-sensing methods. A good primer on RF-based methods for instance is given by Wang et al. [165] where with the Fresnel model the underlying principle of these methods is presented. Recent literature reviews with a more detailed overview of contactless respiration measuring methods in general, and for depth-based methods in special can be found in [115] and [4], respectively.

2.3.1 *RF-Based Methods*

More recently, several works have focused on the detection of breathing rate and non-invasive detection of breathing-related disorders with RF monitoring systems that extract the breathing signal from the wireless channel by taking advantage of the Doppler effect, where the movement of the torso during breathing causes a Doppler frequency shift [40]. The biggest advantages of RF-based methods are that the respiratory rate can even be detected from persons behind obstacles, any kind of lighting is not required, and privacy concerns due to image recording cannot arise. Devices applied here range from Ultra-Wideband Radar, Continuous Wave and Frequency-Modulated Continous Wave, up to even standard commodity WiFi devices. UbiBreathe [3] for instance presents an approach that works on WiFi-enabled devices, even when the device is not held to the chest by the user. Evaluations on three study participants have shown that under certain settings such an approach works well, but is heavily influenced by user's motion and on the location of the wireless access point and the wireless device. Furthermore, the TensorBeat system [166] employed CSI phase difference data to obtain the periodic signals from the movements of multiple breathing chests by leveraging tensor decomposition. Their work shows in a larger-scale experiment in multiple environments that breathing rate estimation becomes particularly challenging when more people are present in the environment. Wang et al. [165] derived with the Fresnel model the underlying physical principle for RF-based respiration monitoring. In their work, it is shown how WLAN based respiratory rate detection depends on location and orientation towards receiver and transmitter, and how a two user respiratory rate detection under ideal conditions is challenging. Both users need to breath at a different pace to be able to distinguish the signals and it is not possible to assign a signal to the respective person. The location dependency was leveraged by [176] through conjugate multiplication of CSI between two antennas. The biggest challenges of RF-based respiration estimation, as pointed out in [40] and [138] are: Problems that arise with the multipath effect, motion artifacts corrupting the Doppler shift on the torso movement while

breathing, interference with other medical equipment, and high power demands of some techniques while devices that can operate with less power, i.e. WiFi, generally have a lower sensitivity. For these reasons, highly precise systems are very complex and costly.

2.3.2 *Non-Depth-Based Optical Methods*

Respiration estimation has been proposed using standard RGB and near-infrared cameras early on. These optical methods most commonly compute optical flow, e.g. using Lucas-Kanade [101] or Horn-Schunck [63] methods, to extract the respiration signal from a video stream, such as techniques presented in [122], [121], and [90], but also approaches using image subtraction techniques exist, such as [156]. In Bauer et al. [14], the respiratory rate is measured with both, optical flow computation with the combined local-global method [23] and a depth sensor with surface registration as proposed in [13]. In this latter paper, the respiration measurement based on optical flow delivered a more accurate respiratory rate estimate compared to mere Time of Flight (ToF) depth measurements. This finding can also be supported by [76], which shows that human breathing mainly occurs along the superior-inferior direction. Consequently, other works also make use of the upward and downward movement of the chest induced by respiration [144, 156]. While usually only a standard RGB camera is required for these methods, they typically have high computational demands and require the implementation of complex algorithms. Other works are based on thermal imaging of the face, where a change of facial temperature is induced by respiration [5, 44]. These methods, however, require close distances of maximum 1 meter, need a clear view of the face, and have to deal with head movements.

2.3.3 *Depth-Based Methods*

The measurement principle of depth-based respiration estimation relies on observing the change in distance of the chest or abdomen towards the depth sensor during respiratory cycles. Inhalation increases the torso volume and will bring these regions closer to the depth camera while exhalation will revert this effect. The change of distance for normal breathing typically is in the range of millimeters to a few centimeters, depending on the person and observed body area. Due to the small distance changes caused by breathing, depth-based methods are susceptible to even slight body movements, especially towards the camera. In most of the related work, the observed person therefore needs to keep still by for instance sitting on a chair with back support or by lying down. In the following, the various depth-based respiration estimation methods are coarsely grouped by their approach.

2.3.3.1 *Distance to a Plane*

Early versions of depth-based methods fixed a plane on the chest and the abdomen of a person lying on a horizontal surface and measured the Euclidean distance of the these planes to the supporting surface plane [130, 145]. Over time, the distance changes of these planes reflect the person's breathing movements. In both papers, it

is argued that a ToF camera has certain advantages over other depth systems, such as a higher accuracy, no need for calibration, and it being more suitable for real-time capabilities.

2.3.3.2 *Motion along Front Axis*

A later method by Noonan et al. [125] uses a fixed 10 cm x 20 cm rectangular selection on the center of the person's thorax, where the mean orientation of this rectangle is computed over 10 successive image frames. The motion component along the surface normal then becomes the estimate of the person's respiratory rate.

2.3.3.3 *Volume*

In other works, the volume of the user's chest or torso explicitly is modelled from depth data. The successive works by Aoki et al. [8–10] use the Kinect's shoulder and hip joint position estimates as boundaries for a rectangular selection and convert the included depth values to 3D coordinates. With these coordinates, a so-called quasi-volume of the user's chest then is modelled explicitly by using Delaunay triangulation with linear interpolation. The observed quasi-volume is shown to be proportional to the air volume measured by a spirometer. The method was evaluated by monitoring 6 male study participants on a bicycle ergometer, pedaling at a constant speed and with the motion artifacts present in the obtained signal. These motion artifacts, due to the known pedaling frequency of about 1 Hz, could subsequently be filtered out with a Fast Fourier Transform (FFT) band-pass filter with a bandwidth of 0.1 Hz to 0.7 Hz. Soleimani et al. [152] compute the respiration signal with a volume based approach as well as by taking the mean of the respective depth values. Both outcomes are compared and it has been shown that the volume-based approach was less accurate while being computationally much more expensive.

Since the depth camera does not see the back of the user, the presented volume-based methods bound the volume at a certain constant distance threshold to the back and compute the volume by integrating over the distances of the single surface vertices to this back boundary. In other words, the volume basically is computed with a weighted sum and, apart from subtracting the distance threshold, can be approximated by the mean of the respective depth pixel values. Due to the lower computational complexity, the majority of the proposed respiration estimation methods thus are based on computing the mean, as will be shown in Section 2.3.3.5.

2.3.3.4 *PCA-Based Methods*

To obtain more reliable estimates, previous work has also suggested to explicitly model respiration using Principal Component Analysis (PCA). The PCA model is acquired from a certain number of successive depth images of a predefined area of the user's torso. Wasza et al. [167] for example compute a PCA model of the user's torso and apply the varimax rotation such that the obtained model has more relevance to respiration than the model from the standard PCA. Its principal axes were found to feature local deformations that are highly correlated to thoracic and abdominal breathing, respectively. This work was extended later on by Wasza et al. [168] by integrating multiple depth cameras to yield a PCA-based shape motion model of the

observed person using prior knowledge of the 4-D shape deformation. In this work, also some issues with the varimax rotation are addressed. In [114] a zoom lens is attached to the infrared projector of the Kinect v1 in order to increase the size of its projected light dots. The trajectories of these dots with a length of 30 seconds are stored in a matrix and a PCA is applied to it. With the iterative EM algorithm the 16 strongest components are calculated and all bases that fail the Durbin-Watson-test are thrown away. Furthermore, all bases with less power in the interest region of 0.02 Hz to 1 Hz are discarded and, to reduce noise, an average of the remaining bases is computed. The approach is used for measuring the respiratory rate of sleeping subjects and, while a special Region of Interest (ROI) is not required, only one person can be within the field of view of the Kinect. The approach was tested on 9 sleeping study participants which were positioned in an optimal view and at different distances. It works best at 200 cm. A common method is to place fiducial markers on the chest and abdomen to define the regions that are used to extract the depth measurements from. Wijenayake et al. [171] for instance use white markers visible in the RGB data of an Asus Xtion PRO RGB-D camera and compute a PCA model from the first 100 depth frames by only using the depth readings inside the region defined by the markers. The first three principal components of such a patient-specific model then are used to reconstruct a noise-free surface mesh. The change of volume of such a mesh has shown strong correlation to spirometer data. For this model, the frames have to be preprocessed to reduce noise and to fill holes, as the PCA is very susceptible to it.

2.3.3.5 *Mean-Based Methods*

A simple proof of concept of measuring the respiratory rate with a Kinect v1 structured light depth sensor is presented by Xia et al. [173]. Here, a solid plane is attached on the chest of the examined body. It defines the ROI and acts as a translation surface. The depth values of the plane's surface points are averaged for each received depth frame and thus reflect the average distance of the plane to the Kinect sensor at the different time instants. The key idea is, that the chest elevation during breathing is expected to cause most depth pixels, and thus the average among all pixels, to correlate with the breathing motion.

In [27] and [146], a Kinect v2 is used to observe the respiration of sleeping persons and classify different sleep states (being awake, in REM, or non-REM) by using features that contain the frequency and the regularity of the breathing. The respiration signal is obtained from the average of the depth values within the hand-annotated chest region. Furthermore, in [146] also the averages of pixel-wise depth differences over two successive depth frames are computed, and [27] applies linear interpolation between two successive depth frames to by-pass non-equidistant sampling caused by the depth camera and uses a wavelet transform to de-noise the results. In [160] an error of only 0.21 breaths/min is reported on a method based on computing the mean of the depth values within a target area.

Benetazzo et al. [15] use the shoulder and hip joints as delivered by the Kinect v2 SDK to determine the region of interest. All depth values within that region are averaged per frame, followed by a weighted average of four successive mean values to reflect the respiration data over time. This work is the first to test different param-

eters for a mean-based approach. It includes sampling rates being varied between 5 Hz, 7 Hz, and 9 Hz, different orientations (0° or 25°), three different light intensities, and variable clothing worn by the observed person (sweater, jacket, and T-Shirt). The evaluation however is approach-specific and instead of a detailed parameter evaluation, results only show that the parameters tested have in the end little effect on the proposed algorithm's performance.

With the addition of RGB data that is available in many depth cameras, extra biophysical information can be extracted. Procházka et al. [137] in addition to the respiratory rate also estimate the heart rate by using the Kinect's built-in RGB and infrared camera to detect the slight changes in color around the mouth caused by blood pressure changes for each heart beat. The respiratory rate is, as in previous works, obtained by averaging all depth pixels within a rectangular selection at the torso. Both signals are band-pass-filtered with the respective cut-off frequencies (0.2 Hz and 2.0 Hz) set in such a way that the frequency components that are not part of breathing or the heart rate are rejected.

2.3.4 *Applications of Remote Respiration Estimation*

This section has the aim to present some potential applications of remote respiration estimation. The focus hereby lies on two topics: Human activity recognition and health care. Since most methods and studies described above were conducted in the context of a specific application (mostly health care), this section is meant to complete the list of applications. Previously mentioned works thus will not be repeated here.

In the domain of human activity recognition, there exist some studies that have incorporated respiration into their experiments. Interestingly, in many of these studies, either a depth sensor already is used, e.g. to track a person's movements, or a depth camera could easily be deployed, but none of them use it to estimate respiration. In all these studies, a remote sensing of a person's breathing using a depth camera thus could effectively reduce the amount of sensors a person has to wear.

Centinela [92] uses acceleration data in combination with vital signs, including respiratory rate, to distinguish different activities, namely walking, running, sitting, ascending, and descending. It is reported that vital signs together with acceleration data can be useful for recognizing certain human activities more accurately than by considering acceleration data only, especially in the case when acceleration signals are similar. The classification of some activities on the other hand did not benefit from the additional vital sign data. In [26], physiological data, including respiratory rate, obtained from a wearable sensing device is used as auxiliary modality to discriminate between four activity classes, namely lie, sit, walk, and jog. To recognize lifestyle activities of diabetic patients [104], WiFi, GPS, sound and acceleration data from a smartphone, as well as heart rate and respiratory rate from an ECG monitor are used to distinguish ten different classes. The effects of respiratory rate on classification accuracy for the latter two studies, however, was not evaluated.

An interesting application is Go-with-the-Flow [151], where body and movement awareness is enhanced through sound feedback to rebuild confidence in physical activity for patients with chronic pain. The sound feedback is generated based on a patient's movement and breathing. The patient's posture, inter alia, is tracked via a

Kinect sensor whereas breathing is assessed with two wearable respiration sensors. Because breathing rate rises with anxiety, and patients often hold their breath if they are anxious or overly focused on a movement, the system, based on evaluating the respiration data, produces sound signals as a prompt to breathe calmly.

In terms of medical prevention and rehabilitation, there exists a dataset of functional senior fitness tests [18] that comprises acceleration readings at the hip and posture data from the Kinect, as well as ECG, respiratory rate, and blood volume pressure from physiological sensors. This dataset is meant to develop algorithms to automate the assessment of fitness levels of seniors. Other applications of remote respiration estimation that are specific to medical care for instance are studied in sleep laboratories. Here, depth cameras have been used to identify sleep apnea in patients that are monitored remotely while lying in a bed [6, 147]. A subject's respiration signal can also be used for drowsiness detection, for instance while driving a car [56, 65]. Finally, emotion classification from respiration and other physiological features has been a focus in some studies [57, 60]. In [57], an accuracy of emotion classification of 75% to 90% from breathing alone is reported, depending on the chosen feature within the respiration signal.

2.4 SUMMARY

This chapter compiled the most recent methods and state of the art of the three main topics *Complementary Motion Sensing*, *Piecewise Linear Approximation*, and *Remote Respiration Estimation*. In the following, a short summary as well as the relation of this work to the available literature will be given.

To achieve complementary motion sensing, there already exists a broad variety of algorithms to fuse inertial and optical motion data as discussed in Section 2.1. Successful applications, however, are not widely seen although promising results can be expected from fusing both. Exact reasons remain unknown, but it is likely that the potential limitations inherited from both modalities are discouraging while the benefits of such an approach are not that obvious. For this reason, during this work it will be investigated how inertial and optical motion data can be used in a complementary way, with the aim to show the benefits of such an approach and to clear the path towards an efficient utilization of both modalities in combination. A method towards such an efficient utilization would be the automatic identification of the person and limb each single IMU-equipped sensing device is worn on within the pose data obtained from the depth sensing. The literature hereby provides only few works that did something coarsely related to that. Methods that did something similar still require the IMU to be assigned manually to the person wearing it and overall are not designed to identify a person or limb in the first place, but try to track moving objects or people in the image frame [62, 72]. To close this gap in this research, a novel method is proposed that allows to efficiently identify the person and limb an IMU is worn on by comparing its motion data to all pose estimates as obtained over time from a depth camera that inter alia observes the respective person. This method is not limited to such a use-case, but potentially can open up a wide range of applications that use complementary motion sensing to achieve more than just sensing motion, as will be detailed in Section 3.

To reduce the amount of motion data that needs to be transmitted, PLA has been found a viable compression scheme that has previously been applied successfully to compress various types of data while keeping important trajectory information of the compressed signal. PLA algorithms, however, have so far never been used for the approximation of quaternion-based orientation sensor signals. To close this gap in the research, Section 2.2 provided a first analysis and comparison of existing PLA algorithms with respect to many constraining factors that limit their ability of approximating quaternion-based sensor signals (see Table 2.1). A focus hereby lies on their applicability in environments with limited processing and memory resources, as can typically be found on wearable devices or stand-alone sensor nodes. Following this comparison, the specific requirements for compressing unit quaternions with PLA techniques will be discussed in Section 4.2. This ultimately will lead to fastSW, a novel PLA algorithm that unifies the advantages of state-of-the-art PLA methods with respect to memory consumption and execution time, but also the choice of segment points and approximation quality as required for the compression of unit quaternions.

When it comes to remote respiration estimation, it could be seen in Section 2.3 that all methods based on using depth data to obtain a respiration signal are only evaluated on a few study participants that furthermore were specifically asked to either lie down in supine position or to sit still in a chair. In some studies, they even had to wear tight clothing. In fact, there is no study where participants are allowed to stand freely or even to occlude their upper body, and in the rare examples where participants are allowed to move, this movement still is heavily constrained, for instance to pedal with a constant rate of 1 Hz on a bicycle ergometer such that this motion component can easily be filtered out later on [9]. Also, there is only a single study that is evaluated on different parameters, i.e. three sampling rates of 5, 7, and 9 Hz, different user orientations from 0° to 25°, and different clothing (sweater, jacket, and T-Shirt) [15]. In the end, this study, however, only states that the parameters had little effect on the results without elaborating on these parameters in more detail. Furthermore, it is difficult to compare these methods among each other or to assess their performance in more realistic settings since they are all evaluated under their own specific conditions and parameters that cannot easily be extrapolated to other methods. A publicly available benchmark dataset as well as a systematic evaluation of important parameters that influence the various respiration estimation methods is missing. To overcome the limitations of previous methods, a novel method will be proposed that in contrast to most of the above approaches does allow users to stand upright and even move their arms and hands in front of their torso. Also, two datasets will be recorded and made publicly available, one for the validation of novel depth-based respiration estimation methods and one benchmark dataset that allows to assess the performance of different depth-based respiration estimation methods under a variety of different parameters. Finally, another issue that will be addressed in this work is the lack of a systematic evaluation of important parameters that influence the various respiration estimation methods as well as a comparison of these methods per se.

3

COMPLEMENTARY MOTION SENSING

This chapter is based on the peer reviewed publications [77] and [82]. Some passages have been quoted verbatim. I am the first author of both publications.

3.1 INTRODUCTION

A variety of different techniques to capture human body motion has been proposed in the past, ranging from different types of sensors worn on the body up to remote sensors in the environment that all in some way measure limb orientations or joint positions. The most prominent solutions comprise either body-worn inertial sensors, typically in the form of a sensor equipped body suit [143], or a number of highly specialized cameras in the environment that make up an optical Motion Capturing (MoCap) system [1, 2]. Optical MoCap systems hereby typically are the most accurate systems, but also tend to be very expensive. More recently, also solutions were proposed that, with the help of machine learning, can estimate human pose from a single depth [150] or RGB camera [25, 102, 116, 117]. These methods come with little to no extra costs in terms of equipment, but at the expense of reduced accuracy. When using on-body sensing methods, in many cases only the tracking of certain limbs, e.g. the arms or legs, is of interest, but more recently, also techniques have been proposed that only require a sparse setup of body-worn Inertial Measurement Units (IMUs) to capture the full body [64, 119].

What all MoCap techniques have in common is that they all come with their specific strengths and limitations that are inherent to the used modality. On-body sensing is cheap and allows for high mobility and continuous monitoring, but is less accurate, may be constraining to wear, e.g. when the full body is captured, and often is limited to particular limbs. Remote sensing from the environment, on the other hand, often is more accurate (e.g. if a high-end optical system is used) and the observation of the full body of multiple persons is possible with the same sensor setup. It is however limited to a certain working volume, in most cases not so easy to set up, and it is sensitive to occlusions. Interestingly, both modalities seem to complement each other very well. The limitations of one modality can be compensated by the other and vice versa. A complementary use of both modalities has already been taken up in prior works and nowadays, as discussed in the literature review, there exists a variety of algorithms to fuse inertial and optical motion data (also see Section 2.1). Successful applications, however, are not widely seen, although promising results can be expected from combining both modalities, with examples ranging from health care and medical applications over sports up to activity recognition or even ordinary daily applications. One reason might be that also the limitations of both systems are inherited and that the steps towards an efficient utilization of the combined modalities remain unclear or simply are not straightforward to achieve.

For this reason, in the following it will be investigated how optical motion capturing and body-worn, inertial motion capturing can be used in a complementary way. Accordingly, a method is proposed that aids in combining both modalities by matching inertial and optical orientation data such that the person and limb the inertial sensing device is worn on can be identified when being observed by a camera. This method does not only simplify the process to set up an inertial MoCap system by automatically affiliating each sensor with the respective person and body part it is worn on, but it also allows a wearable device and an external camera-equipped system to identify each other and subsequently to communicate user specific data. Thus, it can be considered an important algorithmic component for a range of interesting applications. The wearable can for instance benefit from data like the user's indoor location or its full body pose, e.g. in fitness applications, and the camera system can profit from being able to reidentify persons after leaving and again entering the field of view, for instance to track a security person carrying valuable goods. Moreover, such services depend on the wearable device and can at any time be turned on or off, e.g. due to privacy concerns. These examples nicely demonstrate how the complementary use of different MoCap modalities can lead to applications that achieve more than just capturing motion.

To reach that goal, first an overview over technical considerations for combining multi-modal sensor input within a common system is given. This includes a detailed discussion of the different reference frames of the respective modalities, as well as calibration considerations to join different modalities into a common frame of reference, thus enabling a collaborative work between them. In a short preliminary case study, furthermore, the characteristics of a simple approach of replacing limb observations from depth based MoCap with those obtained from IMUs are assessed in a user study. The goal hereby is to investigate the peculiarities of both modalities and to assess if it is feasible to use optical MoCap data to represent the body while using IMU data to represent the dominant arm and wrist. Finally, the method for matching optical and inertial motion data is introduced and evaluated in a user study.

3.2 HANDLING MULTI-MODAL SENSOR INPUT FOR MOTION CAPTURING

Working with data from different sensor modalities and integrating them into a single system is due to the different nature of the sensors and their data not trivial and requires some considerations. In the following, a list of the most important considerations is compiled that need to be paid attention to each time a new sensor type has to be integrated into a common system. There are no sharp edges between the different considerations and some of them might be less important for a specific sensor system than others, but in summary they reflect the creation of a common base on which the sensors can communicate with the system and can agree upon the measured data. Moreover, for most of these considerations, there is no clear answer or path on how to achieve good results, since this heavily depends on the used sensors, the overall system design, and the desired results. Thus every system has to be evaluated on its own.

3.2.1 *Technical Considerations for Combining Different Modalities*

Sensors and Modalities. Among the first steps is the identification and selection of the required sensors and sensor modalities as well as the assessment of their properties and working principles. This includes knowledge of the sensors' limitations, capabilities, and requirements. Some sensors need a specific sensor setup or do not work well in some environments, such as under bad lighting conditions in the case of an optical system or the susceptibility of IMUs to magnetic field distortions in proximity to ferrous objects. Also other sensor peculiarities have to be accounted for, for instance different sampling rates, different reference frames, or missing data e.g. due to occlusion or lost data packets on the used network.

Data and Data Acquisition. Different sensors do measure different properties and provide different data and data types. Examples are joint positions, joint angles, sensor or limb orientations, magnetic field, acceleration, or angular velocity. Different sensors might capture a single limb, the whole body, or even multiple persons. Knowledge of each sensor's data and its meaning thus is imperative. After receiving, the data might need to be converted to a common scale, common unit, or common reference system. Data transmission can happen in two general ways, wireless or with wire, e.g. via Bluetooth, WLAN, or USB. Various data transmission protocols have a different range, data rate, timing, and reliability of connection. For instance, single data packets may get lost, interchanged, or may arbitrarily be delayed. When establishing a connection to a sensor, this should be under consideration.

Synchronization and Timing Considerations. Different sensors have a different notion of time, i.e. a different start time or reference time, and they have different sampling rates and possibly a varying accuracy of their internal clock. For this reason, the received data has to be synchronized to the common system time in order to allow further processing. If no further information is available, the time of arrival of a data point is the only source of a reliable time stamp. Generating and transmitting a time stamp directly on the sensor for each individual data point, however, does aid in synchronization, e.g. in case of delayed or interchanged data.

Sensor to Segment Mapping. In order to be meaningful, the motion data obtained from the sensors somehow needs to be affiliated with an avatar or, more precisely, with the respective bones of an avatar's rig. This affiliation is called sensor to segment mapping and usually has to be done manually. Depending on the underlying sensor system, the sensor data can either be captured from a single limb, such as in the case of an IMU, or it can comprise motion data from multiple body parts or even from multiple persons at once, for instance when using an optical MoCap system. This makes a sensor to segment mapping hard, since such a mapping is not straightforward to achieve when a sensor represents more than a single limb. Furthermore, the data of some body parts might be discarded or be affiliated with a different or even with multiple different segments. A single segment also might receive data from multiple sensors

after being fused in a previous step, especially in the case when the same limb is observed by different sensors. An easy to implement and flexible way to deal with all these challenges is to create a virtual sensor node for each sensor and observed body part. The motion data of each single sensor then only needs to be split up into the different body parts and be distributed to the corresponding virtual sensor nodes. These in turn can be affiliated with one or more segments. Each virtual sensor node thus is linked to a certain real sensor and the respective body part the data comes from, but at the same time acts as if it was a stand-alone sensor that only observes a single limb. The advantage of this approach is that all sensor data ultimately is treated the same, effectively abstracting away the underlying sensor system.

Sensor Model. Different types of sensors might require a specific sensor model to compensate for sensor specific characteristics. These include intrinsic as well as extrinsic parameters that need to be determined via a sensor specific calibration procedure. Optical sensors for instance might need to model intrinsic camera parameters, e.g. to correct lens distortion using a camera matrix [170], and IMUs need to maintain calibration parameters to correct for inaccuracies of the internal accelerometer, gyroscope, and magnetometer. On-body sensors, furthermore, are placed on a person's body surface, i.e. on its skin or clothing. The body surface neither is flat nor is a sensor always placed on the same location or in the same direction and thus, a certain offset to the true limb orientation needs to be modelled. Furthermore, on-body sensors are subject to soft tissue deformation [36], might loosely be attached, or experience sensor drift.

Calibration. Calibration is the process of finding a sensor's intrinsic and extrinsic parameters that correct for errors in its measurement procedure and that map its data to a common scale and reference system. Calibration is typically a non-trivial process that needs to be performed separately for every sensor and sensor type in use. Many sensors do already provide factory-based calibration parameters from the manufacturer or do an automatic calibration during start-up for intrinsic parameters. Extrinsic parameters, however, need to be assessed, such as the offset between different reference frames or a sensor's alignment error on the body surface, the so called IMU sensor-to-segment misalignment [46]. Calibration should happen from sensor level to system level, i.e. starting with intrinsic parameters such as assessing magnetic field distortions or camera matrices, followed by sensor specific extrinsic parameters such as the sensor-to-segment misalignment, up to higher level calibration considerations such as matching the sensors' reference frames by determining their individual offsets to a global reference frame.

Data Processing. Data processing refers to the transformation of raw sensor input data into meaningful data for the system. Meaningful data in this case is motion data that describes an avatar's joint positions and orientations. Data processing can happen in multiple stages after receiving raw sensor data and can, inter alia, include filtering, interpolation, sensor fusion, or any type of transformation. Examples are the removal of noise from a signal, interpolation to match a target frame rate or to approximate missing data, fusing sensor data

from accelerometer, gyroscope, and magnetometer readings into orientation data, transforming a rotation matrix into an unit quaternion, or utilizing body constraints and interpolation techniques in case of sparse or missing sensor data or low body coverage to approximate missing data along the kinematic chain. Also forward or inverse kinematics can play a role in order to deduce joint positions from orientations or vice versa.

Resampling Motion Data. At some point, there is ready-to-use motion data coming from different modalities and data streams and it is required to bring it together, for instance to display it at a certain frame rate or for recording it into a motion file with a desired target frame rate. The data streams, however, might potentially comprise different sampling rates, e.g. when sensors with a low sampling rate are used in combination with fast sensors. Also, some sensors might sample their data irregularly or data points occasionally are missing due to occlusion events on optical sensors or due to an unreliable data connection. In such cases the data needs to be resampled in order to approximate missing data or to achieve a constant target frame rate. To avoid introducing deviations early in the processing chain, the data moreover should be resampled as late as possible to ensure all (pre-)processing steps are performed on the original, undistorted data at an appropriate frame rate. To approximate a value y_k at time t_k between two successive values y_{n-1} and y_n at time points t_{n-1} and t_n, respectively, an interpolation function Interpolate(x_1, x_2, t) can be used as:

$$\begin{aligned} y_k = \text{Interpolate}(y_{n-1}, y_n, \frac{t_k - t_{n-1}}{t_n - t_{n-1}}) \\ t_{n-1} \leqslant t_k < t_n \qquad k, n \in \mathbb{Z} \end{aligned} \tag{3.1}$$

Resampling the motion data can be achieved by using Spherical Linear Interpolation (SLERP) on orientation data and similarly by using linear, polynomial, or spline-based interpolation techniques on position data. As a result, a reasonable approximation can be obtained at the respectively specified time points in between two successive motion data points.

In summary, in order to efficiently work with data from different sensor modalities, it is required to abstract away their specific peculiarities to be able to treat them as if they were received from a single sensor type. The abstracted version in this context is a virtual sensor node. It hides all required preprocessing steps and acts as a single sensor that can be attached to a certain bone of the avatar's animation rig. Furthermore, it only delivers data of a single specific limb of a certain observed person. A single real sensor such as an optical MoCap system thus can lead to many different virtual sensor nodes, one for each observed limb or moving body part. Distributing the motion data across many virtual sensor nodes serves the purposes (1) that it can be mapped to arbitrary animation rigs with different bone setups, (2) that the data is treated like every other sensor data and existing motion capture routines can be reused, and (3) that it can be interchanged, compared or fused with all other sensor data. This makes fusing optical MoCap data with ordinary IMU sensor data possible by simply fusing the data of two virtual sensor nodes that are specified to belong to the same body part.

3.2.2 *Reference Frames*

A reference frame describes a sensor's local coordinate system, or in other words, it describes what this sensor assumes to be its front, up, and right directions. A camera typically uses its view plane as reference frame such that the camera's sides become the X-coordinate and the up and down direction becomes the Y-coordinate, respectively. An IMU on the other hand typically uses the Attitude Heading Reference System (AHRS) where the geomagnetic north becomes its heading and the gravity vector becomes its negative Z-axis. Consequently, whenever two different sensors measure the same rotation around the same axis with respect to their local reference frame, this does not necessarily mean that they were rotated around the same axis with respect to the reference system of the observer.

It is possible to transition between the different reference frames with the help of a transformation that describes how to map one reference frame to another by translating, rotating, and scaling the reference frame's coordinates such that they exactly match the coordinates of the target reference frame. Typically, a transformation matrix is used to achieve this, but since there is only a rotational offset between the reference frames, unit quaternions can be used here to switch between them.

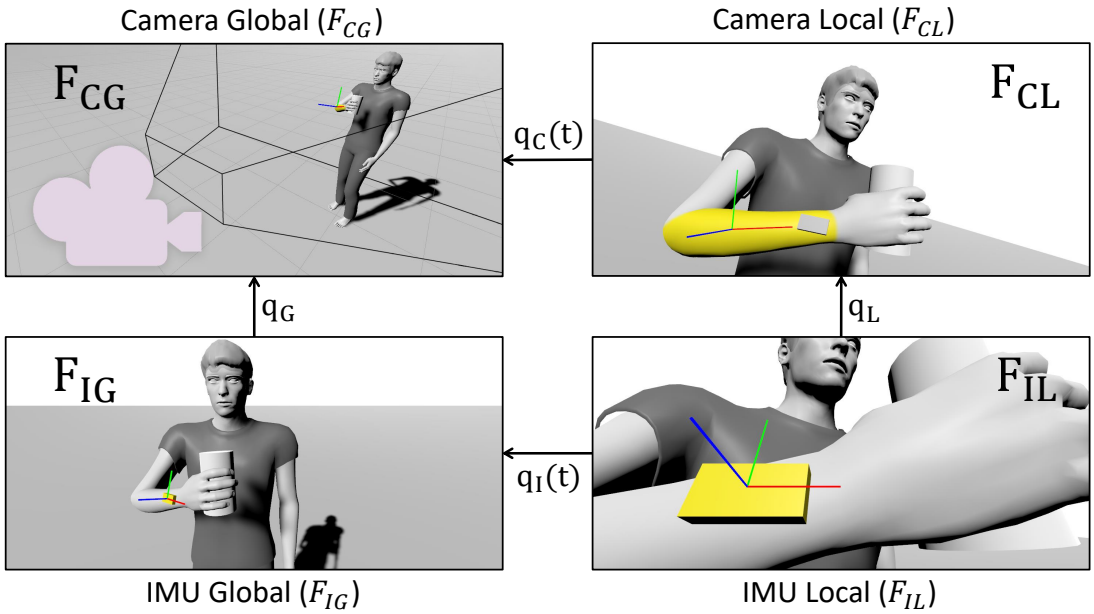

Figure 3.1: From top left to the bottom right: The camera's global and local reference frames F_{CG} and F_{CL} that describe the body segments as seen from the camera, and the IMU's global and local reference frames F_{IG} and F_{IL} that describe the orientation of the IMU. Transitions between these coordinate systems are marked with arrows and can be computed by multiplying with the respective unit quaternion q.

Figure 3.1 illustrates the four different reference frames to be considered, namely the global and local reference frames of the camera F_{CG} and F_{CL}, and the global and local reference frames of an IMU F_{IG} and F_{IL}. The camera frames hereby do not describe the camera's orientation, but the orientation of the respective body segment as seen from the camera. The IMU frames on the other hand do describe the orientation of the respective IMU. Additionally, there exist four distinct transitions between

the reference frames. Transitions $F_{CL} \xrightarrow{q_C(t)} F_{CG}$ and $F_{IL} \xrightarrow{q_I(t)} F_{IG}$ define how to get from the camera or IMU local reference frame to the respective global camera or IMU frame at time t and are obtained as quaternion measurements $q_C(t)$ and $q_I(t)$. In other words, $q_C(t)$ and $q_I(t)$ describe the orientation of the joint or the IMU as perceived from the respective global reference frame, i.e. what is measured by the respective device. The other transitions are defined as $F_{IG} \xrightarrow{q_G} F_{CG}$ to get from the IMU global to the camera global reference frame, and as $F_{IL} \xrightarrow{q_L} F_{CL}$ to get from the IMU local to the camera local frame, respectively. The quaternion q_G denotes the constant orientation offset of the global camera frame to the global IMU frame and accounts for the fact that the camera can arbitrarily be placed in the environment, whereas q_L, also called the segment offset, denotes the rotational alignment offset between the segment as observed by the camera and the IMU that is placed on the same segment. To switch between the different reference frames, the equality of the following transitions can be used:

$$F_{IL} \xrightarrow{q_I(t)} F_{IG} \;=\; F_{IL} \xrightarrow{q_L} F_{CL} \xrightarrow{q_C(t)} F_{CG} \xrightarrow{\overline{q_G}} F_{IG}$$

Expressed as quaternion equation, with $\circ$ denoting the quaternion or Hamilton product, one obtains:

$$q_I(t) = \overline{q_G} \circ q_C(t) \circ q_L \tag{3.2}$$

Note that orientations are chained from right to left, i.e. (3.2) is read as: First rotate by q_L, then by $q_C(t)$, and then by $\overline{q_G}$ (the inverse of q_G).

3.2.3 *Calibration Considerations*

Both, q_G and q_L, usually are not known and need to be determined through calibration, also known as the hand-eye calibration problem in robotics as first described in [149]. To solve the calibration problem, for both $q_I(t)$ and $q_C(t)$, a series of measurements $q(t)$, $q(t+1)$, $\ldots$, $q(t+n)$ is needed. The transition from $q_I(t)$ to $q_I(t+i)$, with $i \in \mathbb{N}^+$, is given as:

$$F_{IL} \xrightarrow{q_I(t)} F_{IG} \xrightarrow{\overline{q_I(t+i)}} F_{IL}$$

Similarly, using the camera measurements $q_C(t)$ and $q_C(t+i)$, there is the path:

$$F_{IL} \xrightarrow{q_L} F_{CL} \xrightarrow{q_C(t)} F_{CG} \xrightarrow{\overline{q_G}} F_{IG} \xrightarrow{q_G} F_{CG} \xrightarrow{\overline{q_C(t+i)}} F_{CL} \xrightarrow{\overline{q_L}} F_{IL}$$

The camera offset q_G cancels out and (3.3) is obtained as:

$$\overline{q_I(t+i)} \circ q_I(t) = \overline{q_L} \circ \overline{q_C(t+i)} \circ q_C(t) \circ q_L \tag{3.3}$$

Substituting $q_A = \overline{q_I(t+i)} \circ q_I(t)$, $q_B = \overline{q_C(t+i)} \circ q_C(t)$, and $q_X = \overline{q_L}$ yields after reordering the so-called calibration equation:

$$q_A \circ q_X = q_X \circ q_B \tag{3.4}$$

Finding the IMU limb offset $q_X = \overline{q_L}$ using a series of n different observations now is subject to:

$$\arg\min_{q_X} \sum_{n \in \mathbb{N}^+} \|q_{A,n} \circ q_X - q_X \circ q_{B,n}\| \tag{3.5}$$

Explicitly solving (3.5) yields for each IMU a specific IMU-to-limb offset estimate q_X, given both orientation data streams match, i.e. are measured from the same limb. An algorithm to solve (3.5) can for instance be found in [12] from which also above calibration scheme was adapted. Solving this equation, however, is a very expensive operation that furthermore is susceptible to noise and unstable orientation estimates. If the matching IMU-limb pairs are not known, it is required to repeat this expensive operation every frame for every IMU for every observed joint in camera space until the respective IMU-limb pairs are identified. Instead, a calibration free matching method should be aimed for, that makes solving (3.5) for all possible IMU-limb pairs obsolete. After matching, and if required, calibration only needs to be performed once per IMU. Such a matching procedure is introduced in Section 3.4.1.

3.3 CASE STUDY ON COMBINING INERTIAL AND OPTICAL MOTION DATA

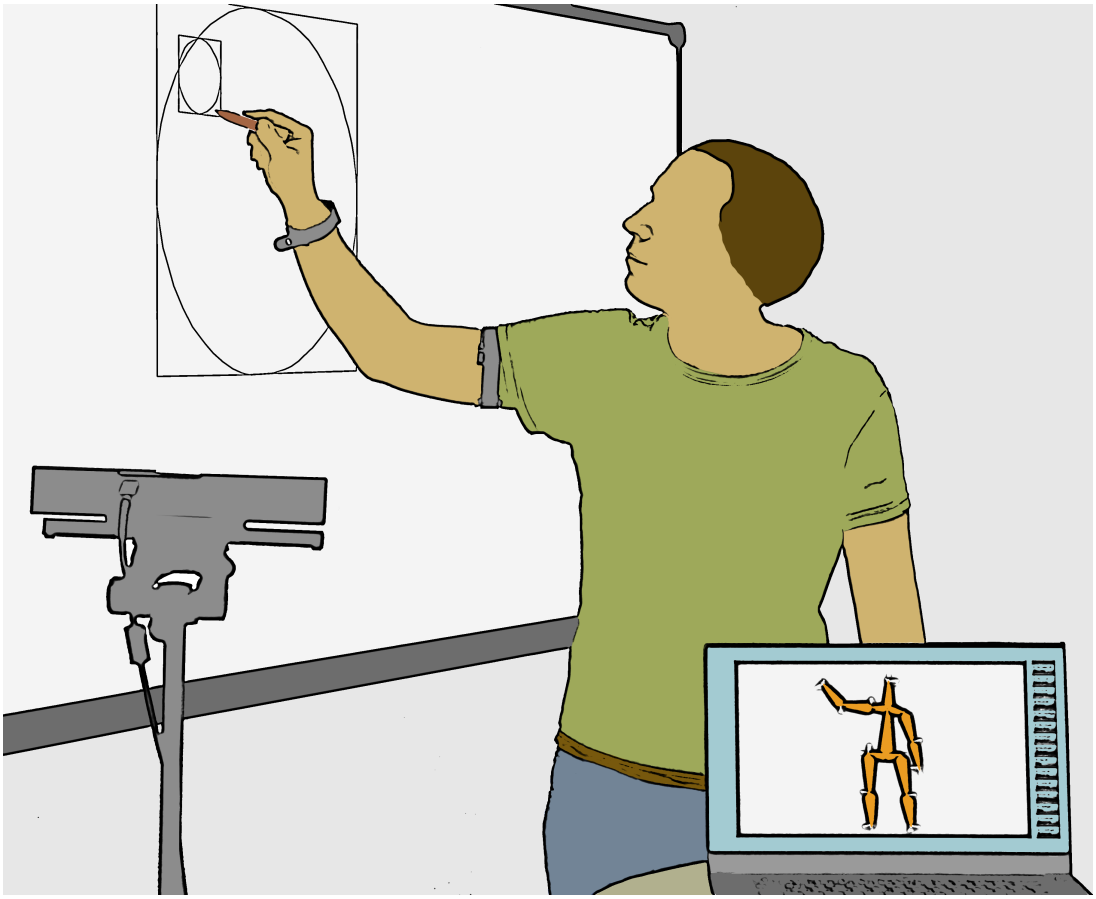

Figure 3.2: The case study combines a Kinect v2 depth camera's posture data with inertial orientation data from two smartwatches worn on the wrist and the upper arm, respectively. The Kinect is placed to the side of a whiteboard on which a user traces a pattern while facing the depth camera.

In the outset of this chapter, it is argued that optical and body-worn sensor modalities can complement each other when used in combination, so the aim of this experiment is to study the effect of combining depth-based and inertial motion data. The experiment is conducted by combining two IMUs, each worn on the wrist and the upper arm, respectively, with the depth-based body posture estimates of the entire person, as illustrated in Figure 3.2. The idea behind this approach is that tracking a person's wrist's position and orientation is a key feature in many applications such as virtual reality, medical applications, computer games, activity recognition, or manual task analysis [181]. For such purposes, the dominant arm arguably needs to be tracked accurately. Combining both modalities is expected to lead to a more accurate system that can cope with common problems that the individual sensors suffer from, in particular occlusions and inertial sensor drift. Furthermore, the IMU can be enhanced with positional information and the human body posture estimated from depth data can be enhanced with orientation data. To this end, this study focuses on how accurate depth imaging and inertial sensing can track a person's hand position. A user

study is conducted under challenging, yet realistic conditions where the observed person is tracing a pattern that is drawn on a whiteboard, potentially causing self occlusions. The pattern itself will be used as ground truth data.

3.3.1 *Study Design*

3.3.1.1 *Sensor Setup*

The Kinect v2 is used as an optical MoCap system. It is a depth camera that primarily is used as a low-cost, consumer grade MoCap system and directly provides joint positions and orientations from all observed persons via the Kinect for Windows SDK 2.0, from which only the joint orientations are kept for the experiment.

The body-worn sensors are represented by two smartwatches, one attached to the wrist and one to the upper arm of the participant's dominant arm. They contain an IMU for sensing their orientation and also have the necessary communication interfaces for wireless data transmission. In the current setup, a custom App lets the Android operation system estimate the orientation of the smartwatches on the device and relay their data via Bluetooth to a nearby smartphone, which in turn forwards all data via a User Datagram Protocol (UDP) broadcast directly to a connected PC.

To enable the possibility to interchange data of the Kinect joints by data from the IMU sensors, for each Kinect joint and each smartwatch an own virtual sensor node is created on the software side, as described in Section 3.2. Each joint's orientation data consequently is treated the same, no matter from which sensor it comes.

To simplify calibration, the Kinect's reference frame is set as the global reference frame. The rotational offset of each IMU (of the respective smartwatch) to the global reference frame then is determined manually by arranging them in such a way that the axes of their local reference system align with the global reference system. The measured orientation is stored as offset to enable manual calibration. To further simplify calibration, the IMUs furthermore are placed carefully on the limbs in a way such that they align with the bone orientation as measured by the Kinect and such that they experience as little soft tissue deformation as possible, which in the following then will be neglected.

3.3.1.2 *Dataset and Experiment Setup*

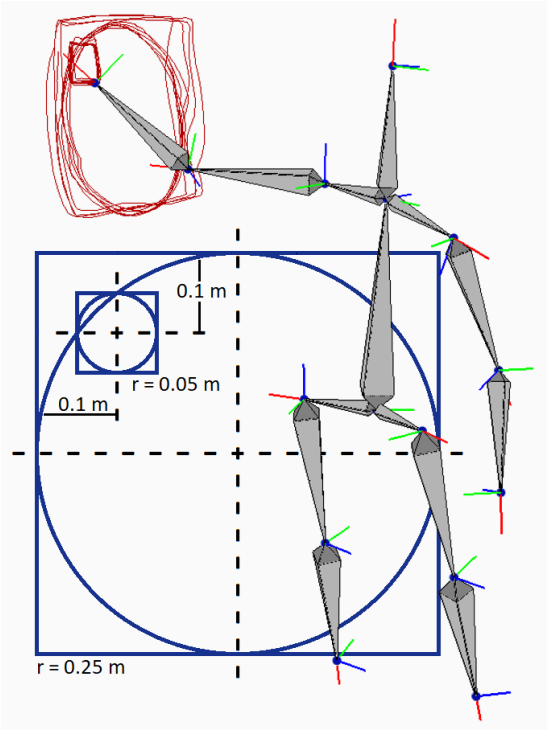

Figure 3.3: The pattern to be traced and the visualization of a human tracing this pattern. The large rectangle and circle have a diameter of 0.5 m and the small ones have diameter of 0.1 m. The real-time trace and captured body visualization, on top of the pattern, provide direct feedback during recording.

For the experiments, 10 study participants were recruited within the University of Siegen. Their body heights were between 1.79 m and 1.98 m tall, and all were right-handed. On a whiteboard, a pattern was drawn that consists of a large and a small rectangle and a large and a small circle (see Figure 3.3). The large rectangle and circle have a diameter of 0.5 m and the small ones a diameter of 0.1 m, respectively. During the motion capture session, the study participants are asked to trace the described pattern with their dominant hand. They were instructed to trace each rectangle and circle at least five times at a pace they could determine. The experiment is conducted in two different settings:

- Setting A: The Kinect was placed to the side of the whiteboard and the current test candidate and the Kinect are oriented such that they face each other during the motion capture (see Figure 3.2). The captured person's front in this setting is fully visible to the Kinect sensor.
- Setting B: The participant was asked to draw the pattern with a natural, self-chosen orientation towards the whiteboard, leaving the Kinect on its previous position to the side. In this setting self-occlusions are not prevented and it can be studied in which extent the Kinect faces problems tracking the arm movements in a more realistic setup.

All participants performed setting A and four out of the ten participants additionally performed setting B.

During the study, the participants' whole body is captured with the Kinect and, as described above, the dominant arm and wrist additionally are captured with two smartwatches. The overall raw sensor data during each motion capture session is recorded such that the session can be restored at any time. This is used to replace parts of the Kinect's sensor data by the respective smartwatches' sensor data from the wrist and upper arm, thereby keeping the rest of the Kinect data in order to provide an anchor to the wrist's and arm's tracking. Overall, three different sensor constellations are tested:

(K): The whole body is captured using only the Kinect v2.

(K + W): The Kinect's wrist capture is replaced by the wrist-worn smartwatch.

(K + W + A): The Kinect's wrist and upper arm captures are replaced by the respective smartwatches worn on the wrist and the upper arm.

3.3.2 *Visual Inspection*

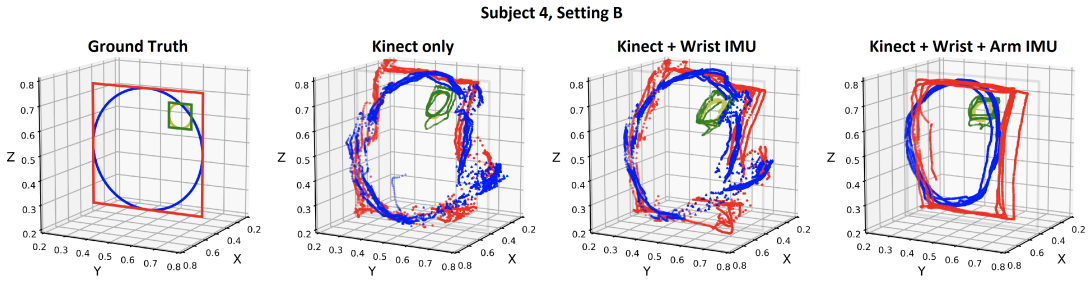

Figure 3.4: The performance of setting B for the three different approaches in case of the user's self-occlusion: The leftmost plot shows the ground truth. The second plot from the left shows the wrist tracking results from just the Kinect's estimates. The third plot shows the Kinect results, with the lower arm segment replaced with the wristwatch's IMU data. The right plot shows the tracking results when both arm segments are replaced by smartwatches.

Study participants were recorded in different settings such that the effect from the level of occlusion can be investigated as a parameter. One of the observations from the first visual inspection of the Kinect's capture data is the detrimental effects that occlusions have on the tracking of the wrist. Only in the very careful placement of the Kinect toward the side of the user (i.e., from the viewpoint of Figure 3.2), the wrist can be tracked at most times with the Kinect alone. Even in such a best-case setup, self-occlusions regularly happen and lead to deviations, as can be seen in Figure 3.5. Also some effects of the smartwatches' IMU drift can be observed: Especially in the X-axis toward the whiteboard, accumulated errors build up in the tracking performance. Overall, from the visual inspection can be concluded that due to occlusion events a continuous trace or a trace without strong deformations cannot easily be achieved by the Kinect alone, but already a single wrist-worn IMU can in many cases improve data quality significantly, despite adding sensor drift.

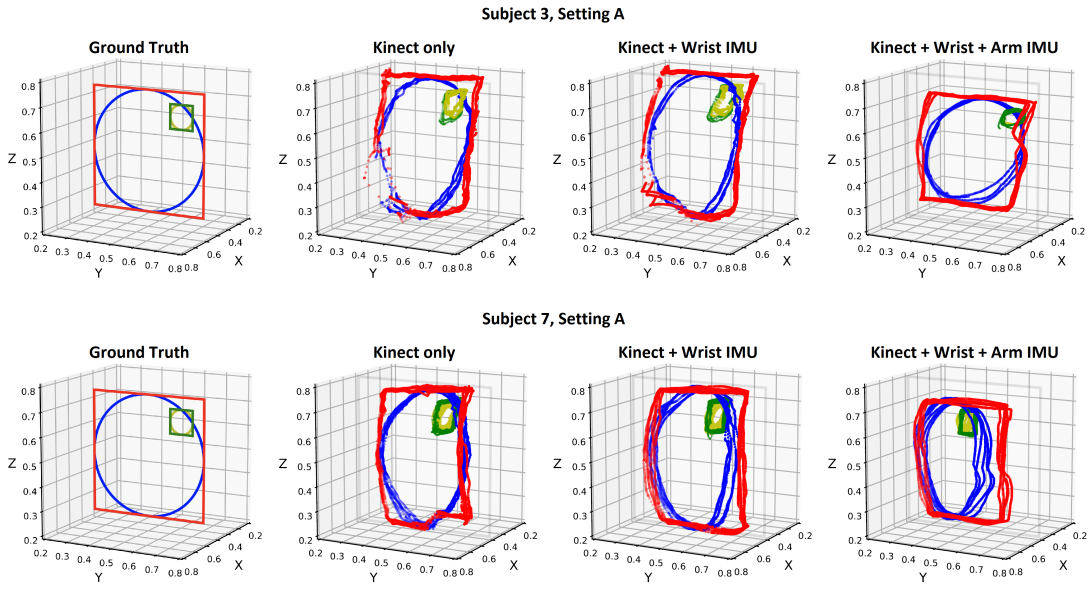

Figure 3.5: Two examples from setting A, illustrating how occlusion distorts the depth imaging's performance as seen on the artifacts in the lower area of the "Kinect only" plots (respective second plots from the left). For these recordings, the Kinect was positioned in a *best-case scenario*, i.e., without occlusion from others and tracked from the participant's side to reduce self-occlusion. Tracking performance is improved by replacing the quaternion for the lower arm with the IMU's data (respective third plots from the left), although the latter contains IMU drift. The drift IMU is emphasized when the upper arm is also tracked by an IMU (rightmost plots).

Due to the aforementioned difficulties to track the wrist with the Kinect alone when major parts of the dominant arm are occluded, as can be seen in Figure 3.4 from setting B, the quantitative analysis is focused on the Kinect's optimal position and a more occlusion-prone sample. This will allow a comparison of the Kinect's best-case performance to track the wrist position, compared to when the upper and lower arm is tracked with a smartwatch.

3.3.3 *Quantitative Analysis*

The quantitative analysis is focused on setting A and thus compares the IMU to the Kinect in a best-case scenario. To assess the performance of the different setups, three quantitative performance measures are introduced: The *Shape Fit*, the *Shape Coverage*, and the *Trace Continuity*. For the first two measures, the euclidean distances between the quantized ground truth points and the recorded trace points of the shapes are calculated.

- The *Shape Fit* is given as the mean and standard deviation of the distances of all single sample points to their respective nearest ground truth points and tells how well the trace points are aligned to the ground truth. Missing trace points or a hole in the trace point pattern (see Figure 3.5), however, cannot be detected due to the distance-to-nearest-point calculation.

- The *Shape Coverage* measure reflects a hole or strong deviation of the sample pattern that always occurs at nearly the same position. It is obtained by computing the mean and the standard deviation of the distances of all single ground truth point to their respective nearest sample point from the recordings. Since only the nearest sample points to the shape are considered, it only reflects how well the best individual trace points are aligned to the shape, not how well the overall trace fits the ground truth pattern.
- The *Trace Continuity* measures the average distance between successive trace points. Since all sensors captured their data on the same recordings and data has a constant sampling rate, i.e. there are no differences in movement speed, this measure indicates how continuous the trace signal is. If a lot of holes or jumps are in the data, the average distance and the standard deviation increase.

Table 3.1: *Shape Fit* and *Shape Coverage* measured as average distance between great rectangle or great circle traces and respective ground truth points, and *Trace Continuity* (TC) measured as average distance between successive trace points.

Setup	Shape Fit [cm]				Shape Coverage [cm]				TC [mm]	
	Rectangle		Circle		Rectangle		Circle		All Data	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
K	4.89	3.36	5.05	3.49	4.53	3.23	4.43	3.17	3.06	4.83
K + W	4.93	3.43	5.15	3.76	3.74	2.90	4.65	3.51	2.81	4.26
K + W + A	6.86	3.03	7.40	4.40	5.99	2.77	7.21	4.15	2.21	1.66

The *Shape Fit*, *Shape Coverage* and *Trace Continuity* (TC) of the overall data set are listed in table 3.1. The Kinect (K) and Kinect + Wrist (K + W) setup perform in terms of the *Shape Fit* with mean values of 4.89cm (K) and 4.93cm (K + W) on the rectangle and 5.05cm (K) and 5.15cm (K + W) on the circle equally well. The wrist sensing thus can easily be replaced by an IMU sensor. The *Shape Coverage* measure furthermore indicates with mean values of 4.53cm (K) and 3.74cm (K + W) on the rectangle and 4.43cm (K) and 4.65cm (K + W) on the circle that at least for the rectangle a better *Shape Coverage* can be achieved by using a wrist-worn IMU. Using an additional IMU worn on the upper arm as in the Kinect + Wrist + Arm (K + W + A) sensor setup introduces errors in the kinematic chain that add up and lead to larger errors on the wrist position. Both, the *Shape Fit* and the *Shape Coverage* are with mean values of 6.86mm and 5.99mm, respectively, much worse than for the other setups. In setting A, where the whole body is seen by the Kinect, the arm mounted IMU therefore does not bring benefits with respect to these measures.

The *Trace Continuity* on the other hand clearly indicates an advantage of using multiple IMU sensors to track the arm. With increasing amount of IMU sensors, the mean values and standard deviations of the *Trace Continuity* successively decrease from 3.06mm and 4.83mm (K) over 2.81mm and 4.26mm (K + W) down to 2.21mm and 1.66mm (K + W + A). This confirms the observation from the visual inspection that due to occlusion events a continuous trace cannot easily be achieved by the Kinect alone.

3.3.4 *Conclusions*

Wrist-worn IMUs are embedded in most smartwatches and can be used to track the wrist's orientation and motion. It is shown how IMU data can improve the capturing of a person's body in a scenario, where the tracking accuracy of the dominant hand is especially important. From the experiment focused on an optimal placement and orientation of the Kinect towards the user (setting A) furthermore follows that a combination of depth imaging and a wrist-worn smartwatch delivers more robust data: The Kinect suffered severely from self-occlusions of the arm when facing the board, and results from where the arm's segments were replaced with IMU data were significantly better, despite minor sensor drift. Especially when a smooth trajectory is required, an additional usage of IMUs to capture the arm can be recommended.

3.4 MATCHING INERTIAL TO OPTICAL MOTION DATA

In this section, a method is presented that allows combining inertial and optical motion data in real-time and without the need of camera-to-IMU calibration, by linking wireless data streams from IMU-based wearables to sets of joints recognized in a camera's field of view, as depicted in Figure 3.6.

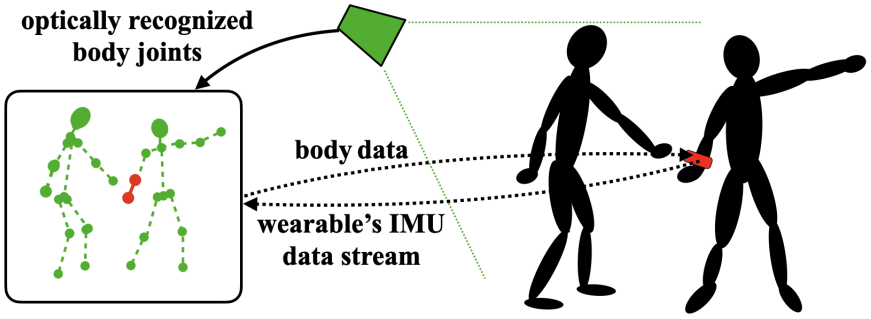

Figure 3.6: The proposed method enables to match wireless streams of IMU data from a wearable device (red) to sets of body joints that have been optically tracked from an environmental camera (green). Once associated, the camera could then send back the user's full body poses to the right wearable as a service.

The idea behind this scenario is that the wireless capabilities of a wearable enable it to stream its orientation, and thus the orientation of the body joint it is attached on, to a camera system in the environment. The observing cameras in turn are able to capture multiple body joints in real-time, provided that those joints are not blocked by surrounding objects, people, or are not out-of-frame. These camera-based systems in the environment could return the full-body joint estimations as well as other valuable information about the user and its surroundings, such as the user's indoor position, back to the user's wearable as a service. As a result, the wearable becomes aware of its user's full body posture and location over time, without requiring users to wear a large number of devices. The body-worn devices on the other hand are not susceptible to occlusion or lighting conditions and can provide valuable data for a seamless body tracking or to resolve ambiguities. For instance, they can be used to enable the reidentification of a user after leaving and again entering the field of

view of a camera. Such complementing modalities can be extremely useful in applications that require full-body tracking, such as virtual reality, character simulation, gesture controlled systems, or activity recognition. In the case of an unknown sensor setup, the proposed method furthermore helps reducing the complexity for setting up such applications as it is able to automatically associate a body-worn device with its respective limb, making a manual assignment of a sensor to the limb it is attached on obsolete. Another promising application is indoor localization, where a person's position, as estimated from a camera system, can be forwarded to a wearable that was identified to be worn by that person.

When it comes to an efficient method to identify the person wearing the inertial sensing device, or detecting the body part that device is worn on, the literature provides only few works that did something coarsely related. Proposed methods try to track moving objects or people by observing and relating their inertial measurement data and their visible movement in the image frame [62, 72]. These methods, however, are not designed to identify the persons by their body-worn IMU in the first place and the IMU still has to be assigned manually to the person wearing it. The proposed method thus closes a gap in the research.

3.4.1 *Method*

The method for locating IMUs in depth video footage consists of three steps. In a first step, the persons and their respective posture needs to be detected. In this case, this is done by the Kinect v2 framework using the method proposed by Shotton et al. [150]. In a second step, the single limbs' orientation estimates from the depth-based pose estimation process need to be compared to the orientation data coming from the IMU with respect to their similarity. In total, four different comparison measures are introduced in the following section. As a last step, the limb-IMU pair that is closest within a certain time frame needs to be identified, taking into account the aforementioned comparison measures.

3.4.1.1 *Quaternion Comparison Measures*

In total, four different comparison measures are introduced in this section that can subsequently be used to match unit quaternions with respect to their similarity. Given the quaternions' bases agree upon the axes of rotation and assuming small camera-IMU and IMU-limb offsets, from (3.2) follows that $q_I \approx q_C$. A first, straightforward option thus is to use the **(1) Quaternion angle** or geodesic angle between two quaternions. It is computed with the quaternion dot product as in (3.6).

$$d_q(q_I, q_C) = 2 \arccos |\langle q_I, q_C \rangle| \tag{3.6}$$

As estimates of the limb joints from camera-based systems tend to be unstable and suffer from randomly swapping around the limb's direction, especially in the case of the forearms, a second measure would be the **(2) Stable quaternion angle**. It makes use of the quaternion swing-twist-decomposition and only keeps the swing

part around the limb direction, in this case the X-axis. The swing quaternion then is forwarded to (3.6) to obtain (3.7):

$$d_{q,stable}(q_I, q_C) = d_q(swing(q_I), swing(q_C)) \tag{3.7}$$

Using the stable quaternion angle also is useful if the IMU's rotational offset around the attached limb is not known or when from the camera-based pose estimation only joint positions are available. The sensor's axis in the direction of the limb, however, has to be known.

Since the assumptions for small reference system offsets made in above metrics do not hold in general, the **(3) Independent quaternion angle** is proposed. Reorganizing equation (3.4) to $q_A = q_X \circ q_B \circ \overline{q_x}$ and considering that the real parts at both sides need to be the same, yields:

$$\begin{aligned} \Re(q_A) &= \Re(q_X \circ q_B \circ \overline{q_x}) \\ &= w_x^2 w_B - w_x \vec{v}_X \cdot \vec{v}_B + w_B \vec{v}_X \cdot \vec{v}_B \\ &\quad + w_B \vec{v}_X \cdot \vec{v_X} + \vec{v}_X \times \vec{v}_B \cdot \vec{v}_X \\ &= w_B (w_X^2 + \vec{v}_X \cdot \vec{v}_X) = w_B = \Re(q_B) \end{aligned} \tag{3.8}$$

From (3.8) follows that the real parts of q_A and q_B are equal, meaning that q_X can safely be discarded. Intuitively, (3.8) can be interpreted as: The amount of rotation or the angle in between two successive measurements at times t and $t+i$ measured by both sensors needs to be the same, independent of the direction or axis of rotation of each. This makes sense since the limb rotation does not depend on the sensor alignment.

With $q_A = q_I(t+i) \circ \overline{q_I(t)}$ and $q_B = q_C(t+i) \circ \overline{q_C(t)}$, both representing the rotation from the respective quaternion $q(t)$ to $q(t+i)$, and considering that both comprise the same amount of rotation in between time points t and $t+i$, the angle between both can be computed using (3.6). The independent distance metric then is defined as:

$$d_{ind}(q_I, q_C) = |d_q(q_I(t), q_I(t+i)) - d_q(q_C(t), q_C(t+i))| \tag{3.9}$$

Considering stability issues of the camera-based limb orientation estimation, similar to the stable quaternion angle, furthermore, the **(4) Independent stable quaternion angle** is introduced. It is based on (3.9), but instead only uses the swing component of a quaternion, as stated in (3.10).

$$d_{ind,stable}(q_I, q_C) = d_{ind}(swing(q_I), swing(q_C)) \tag{3.10}$$

3.4.1.2 *Discrete Joint Matching*

To find the body joint that was picked up by the camera and that matches the wearable's IMU orientation sequence best, first a distance matrix $D[k][n]$ for each camera joint k and sample n is computed using any of the four distance metrics described above. Given a distance matrix D, (3.11) then computes the most likely camera joint k the IMU is attached to at time t and within a window comprising w samples.

$$match(D, t, w) = \arg\min_k \frac{1}{w} \sum_{n=t}^{t+w} \left| D[k][n] \right| \tag{3.11}$$

Equation (3.11) allows to identify the limb position of an IMU at any time point t independently, even if its location on the body or the person wearing it has changed in the meantime.

3.4.2 *Study Design*

To validate the proposed method, a dataset is collected in which three participants are simultaneously captured by a Kinect v2 depth camera while performing near-synchronized movements. One of the study participants was wearing an IMU device that delivered quaternions wirelessly to a system attached to the Kinect, in which also all the participants' joints are calculated from the depth data in real-time. The IMU was worn in two different constellations: (1) on the wrist, as one would wear a smart watch, and (2) in the user's pocket, as one might carry a smartphone. To make the task of estimating on which joint (of overall 42 optically detected joints) the IMU is worn particularly challenging, these three different scenarios with high synchronicity were chosen to evaluate the performance of the methods:

- (A) The *Macarena* line dance, in which participants tend to move one limb at a time, in a synchronous fashion. Participants were at the start of the recording only sightly familiar with the Macarena movements and started asynchronous, though they improved after a few repetitions through listening to and watching the music video as they performed the dance.
- (B) The *head, shoulders, knees and toes* exercise for children, causing participants to move their left and right limbs synchronously. Motion sequences are shorter for this scenario, and participants quickly became familiar with the few movement sequences for this exercise.
- (C) The participants *walking* along the room parallel to the camera's line of sight in a synchronous fashion. In this scenario, the participant with the wearable set the pace whereas the two others were trying to walk in the same pace and rhythm.

For scenarios (A) and (B), the IMU was worn on the right wrist with negligible IMU-limb offset, and for the walking scenario (C), the IMU was worn in the front left pocket with the IMU not being properly aligned to the limb. The camera-IMU offset was about 25° and the IMU was worn by the same participant in all scenarios. Recording times are (A) 95 s, (B) 45 s, and (C) 64 s.

3.4.2.1 *Sensors and Data Preprocessing*

The wearable IMU that is used in this study is a custom wireless sensor module that is built around the Bosch BNO055 IMU, delivering the sensor's orientation as a quaternion at a sampling speed of 100 Hz. It can be used as a single sensor, or combined in a network of multiple IMUs, using Nordic Semiconductor's nRF24L01 low-power transceivers. It runs approximately for 18 hours continuously with a miniature 400 mA battery. For the estimation of the users' joints from the environment, a Kinect v2 framework is used as a well-known depth camera system that performs optical

tracking of users' body joints through a method presented by Shotton et al. [150]. For the datasets generated in the experiments, the detected body joints of all users are stored as unit quaternions at a sampling rate of 30 Hz. The wearable's stream thus is downsampled in order to be able to focus on the matching itself.

To enable the use of the simple quaternion distance metrics, i.e. the quaternion angle and the stable quaternion angle, the quaternions' bases are remapped to a common coordinate system such that they agree upon the axes of rotation. This step is not required for the independent methods, but the mapping usually is known from the used sensing devices and it allows a comparison of the proposed independent methods to the more simple approaches as described in Section 3.4.1.1.

3.4.3 *Evaluation*

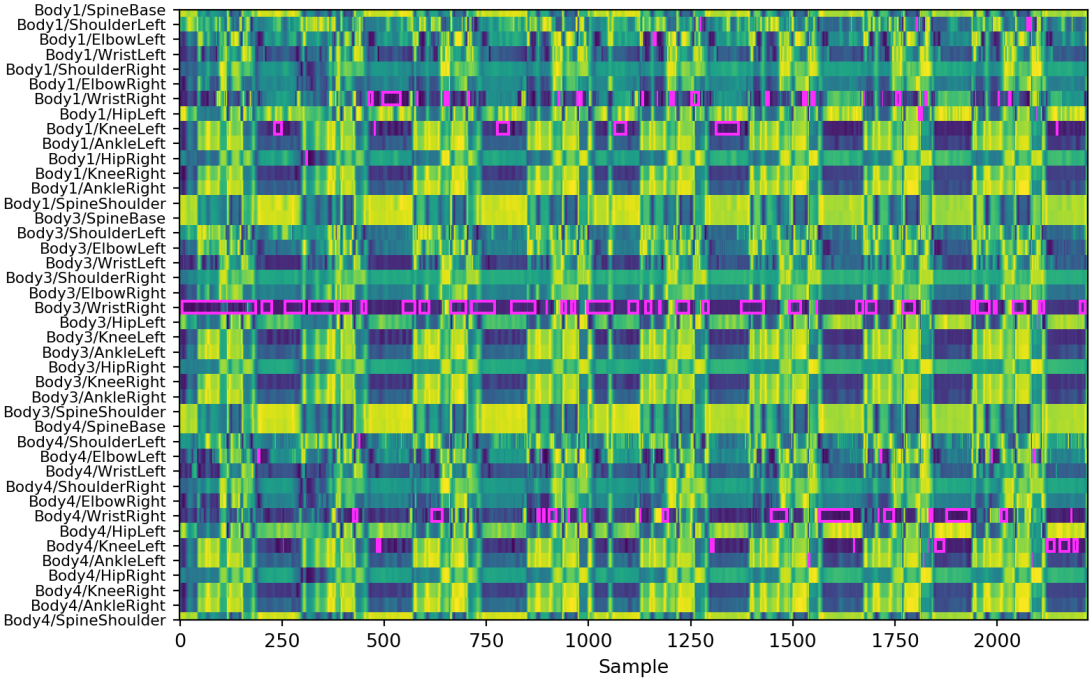

Figure 3.7: Joint distances from stable quaternion distance, equation (3.7), of the Macarena line dance, with evaluated joints in the rows and samples in the columns. Blue being low and yellow being high distances. Joint classification using a window size of $w = 20$ samples is indicated with magenta edges. The IMU was attached to the *Body3/WristRight* joint.

For the evaluation results below, equation (3.11) is evaluated for all distance measures (also see 3.4.1.1). For the independent metrics (3.9) and (3.10), the time offset parameter is set to $i = 7$. At lower values, especially at $i = 1$, the estimation accuracy degrades as there is insufficient movement in between successive samples. An example visualization of the distance matrix from the stable quaternion metric is shown in Figure 3.7. It nicely visualizes the dynamics of the joint distances caused by the rhythm of the Macarena line dance. The IMU hereby was attached to the "Body3/WristRight" joint and the corresponding joint classification is highlighted

in magenta. Although locally other joints have a smaller distance, within broader windows it will overall have the closest distance to the IMU.

3.4.3.1 *Joint Matching Accuracy*

To assess the performance of the different metrics, the window length w step-wise is increased and for each w all samples are classified by moving the window over the respective distance sequence. The joint matching accuracy is computed as the amount of correctly classified window positions divided by the total number of window positions available for a certain w. Figure 3.8 shows the matching accuracy of all three scenarios against an increasing window length w of the moving window.

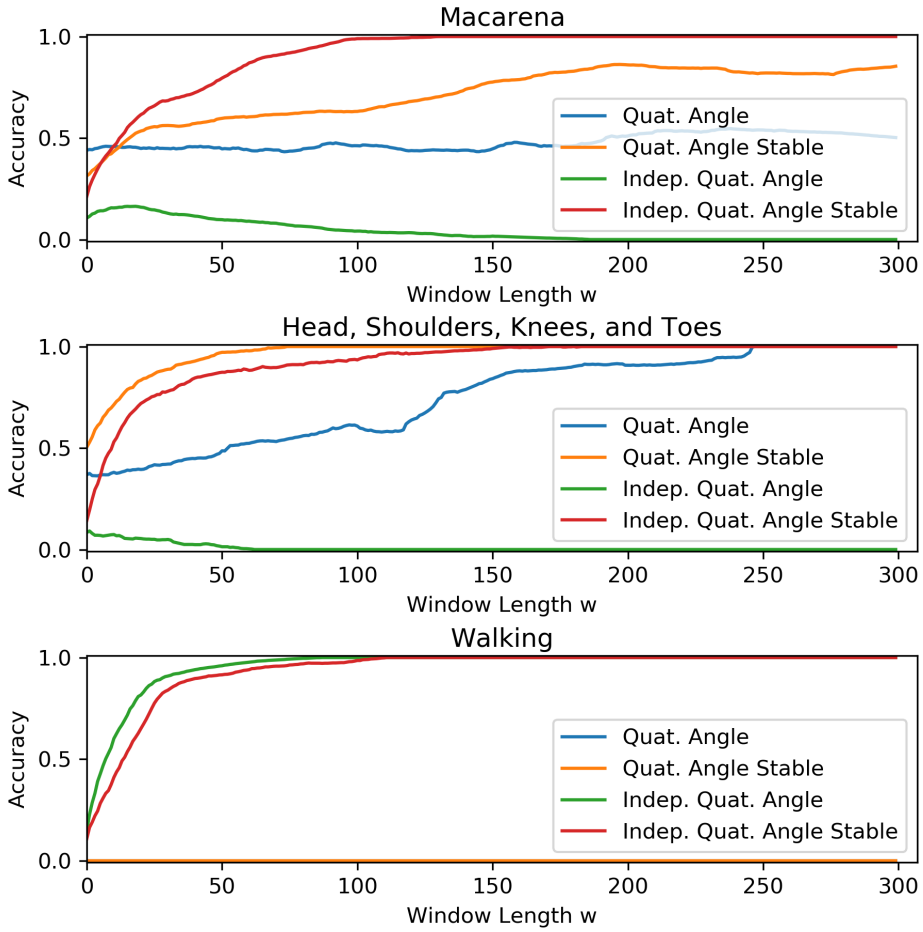

Figure 3.8: Accuracy of the different matching methods, plotted for the different experiment scenarios with increasing window sizes w. The IMU sensor was worn on the wrist in the Macarena and Head, shoulders, knees, and toes scenarios. For walking, it was worn on the hip. The proposed *independent stable quaternion angle* metric can accurately match the correct joint in all scenarios. Depending on the scenario, other metrics nevertheless can show a slightly better performance.

In scenario (A), the stable independent angle distance is the only metric that achieves 100% accuracy, i.e. the correct match can be found at any time. It converges beyond window lengths of $w = 128$ samples (4.27 seconds) and already is close to 100% at $w = 96$ (3.2s). The stable quaternion metric reaches its maximum of 87%

at $w = 195$ (or 6.5s) and the quaternion angle metric stays with small deviations at about 45% over all window lengths. The independent quaternion angle peaks with 17% at $w = 14$ and afterwards decreases to 0%. Since the Kinect v2 has difficulties in correctly estimating the twist orientation of wrist joints, with large angular offsets on successive samples, the methods that are not stable against such errors can not correctly assign the wrist-worn IMU to the matching joint. Especially the independent quaternion angle requires successive samples of both streams to have similar changes in rotation.

A similar behaviour can be observed in scenario (B). The large errors in the camera's wrist orientation estimates cause the independent quaternion angle metric to match other joints that comprise smaller deviations around any axis, resulting in close to 0% or 0% accuracy at all window lengths. The stable independent quaternion angle is able to remove the twist rotation of the wrist and converges to 100% accuracy above a window length of $w = 155$ samples (or 5.2s). The stable quaternion distance metric in this scenario converges fastest to 100% accuracy above window lengths of about $w = 75$ samples (2.5s). The quaternion distance requires at least $w = 246$ samples (or 8.2s) to accurately match the correct joint. In contrast to scenario (A), here both the stable and normal quaternion angle metrics can more efficiently match the correct joint. One reason for this might be that during the Macarena line dance only one limb at a time is moved while during the head, shoulders, knees, and toes exercise many limbs are moved simultaneously and thus any ambiguities can be resolved within smaller time windows.

For the walking scenario (C), only the independent quaternion angle metric and its stable variant are able to correctly assign the IMU to the upper left leg. The first metric hereby converges faster to 100% accuracy at window lengths above $w = 83$ samples (2.8s), closely followed by the stable version at above $w = 108$ samples (3.6s). All other distance metrics can not at all match the IMU to the correct camera joint. The reason is that in this scenario the IMU is not well aligned to the limb and due to this offset any other random camera joint appears to be closer to the sensor orientation at any time.

3.4.3.2 *IMU-to-Camera Offset*

For the case in which the IMU-to-camera offset is not known in advance, the effect is modelled on the assignment accuracy by step-wise increasing the offset from -180° to 180° around the camera's up-axis, with 0° being the unchanged orientation. Figure 3.9 plots the accuracy of all activities against varying IMU-to-camera offsets at a fixed window length of $w = 50$ samples.

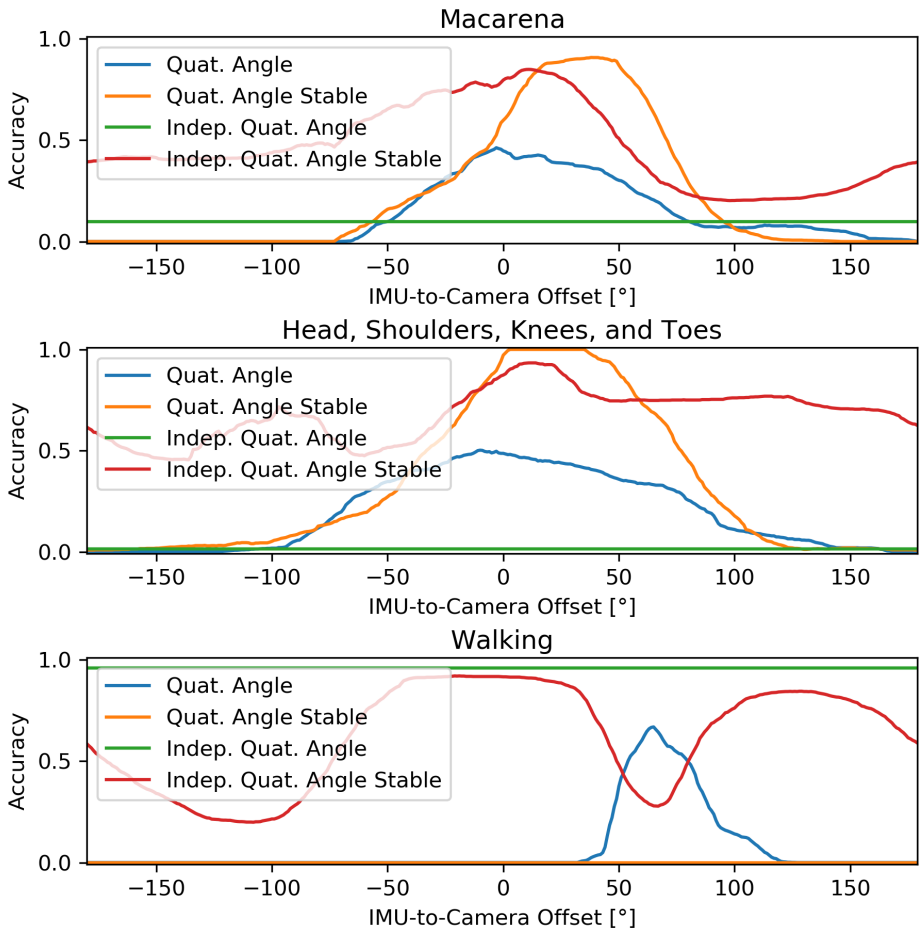

Figure 3.9: Accuracy of the assignment with different IMU-to-camera offsets along the up-axis at a window length of 50 samples, for the three scenarios. Top: Macarena (A); Middle: Head, shoulders, knees, and toes (B); Bottom: Walking (C). For scenarios (A) and (B), the IMU was worn on the wrist, for scenario (C), the IMU was worn on the hip.

In scenario (A), the accuracy of the quaternion angle peaks with 47% at -3°. Its stable variant reaches its maximum of 90% in between 18° and 49°. Both metrics decrease to 0% to both sides. The stable independent quaternion angle shows at 10° an accuracy of 85% and has a decreased performance to both sides, however without dropping to 0%. As this metric relies on the swing-twist decomposition of the measured quaternions, it is not fully independent from the IMU-to-camera offset. The independent quaternion distance is not affected by any rotation offset, but has a low accuracy of about 10% due to the unstable wrist orientation estimates of the Kinect v2.

A similar behaviour can be observed for scenario (B). Here the quaternion angle metric peaks at -10° with an accuracy of about 50%, and the stable quaternion distance reaches its maximum of 100% between 3° to 35°. Both metrics drop to 0% to both sides. The independent quaternion angle again can not deal with the wrist error while its stable variant has its maximum accuracy of about 93% in the range of 7° to 22° and maintains an accuracy of above 45% at all other offsets.

In scenario (C), the quaternion angle metric peaks at 65° with an accuracy of 67% and drops to 0% at both sides. Its stable version remains at 0% for all camera offsets. It is assumed that the quaternion distance metric is at the 65° offset more likely to match the correct joint due to comprising additional rotation information around the X-axis. The independent quaternion angle does not suffer from erroneous camera joint estimates and stays at an accuracy of 96% over the whole range of camera offsets, showing the advantage of truly being independent. Its stable version has its maximum accuracy of 92% between -43° and 25°. Two local minima of about 20% and 27% are between -127° to -98° and at 66°.

3.4.4 *Discussion*

The dataset comprises many relevant challenges of IMU to camera joint matching, namely synchronous movements, erroneous joint orientation estimates, IMU-limb and IMU-camera misalignments, and asynchronous sampling rates. It, however, does not contain a scenario with regular occlusion events and it is with only three different scenarios, three participants and only one IMU per scenario somewhat limited to draw final conclusions about its robustness in real world environments. Especially in case of occlusions, its performance is likely to decrease significantly, but once a joint was successfully associated, a continuous tracking through both complementary modalities is facilitated. Both independent quaternion angle methods do not require calibration, but, in contrast to both other metrics, require user movement for the matching process in all circumstances. The independent quaternion distance metric, however, is susceptible against estimation errors. Its stabilized version can compensate for that, but is, due to the swing-twist decomposition being affected by calibration parameters, less robust against calibration offsets. All proposed metrics only require a few processor instructions and can be computed in parallel for all IMU-camera joint combinations, thus being highly performant even for large numbers of joints. The most important tasks for the future work are to evaluate the methods on a broader dataset, including more sensors as well as occlusion events, and to tackle the dependency of the swing-twist decomposition on calibration parameters.

3.4.5 *Conclusions*

The proposed method allows the quaternion stream from a wearable IMU sensor to be matched, on the fly, with quaternion estimates extracted from an optical sensor (e.g., a depth camera), thus allowing to track the user's full body posture over time. The method accounts for different coordinate systems, as well as inaccuracies that tend to be present in optical body pose estimation frameworks (such as sudden twists in the estimates from the wrists).

A series of experiments was conducted, with participants performing synchronous dance routines, using a 30 Hz depth camera and body-worn IMU sensors. Results show that the proposed method can find the matching joint of the correct user within 75 to 128 samples (or 2.5 to 4.3 seconds) at the wrist, using the stable or the independent stable quaternion metrics respectively. The independent stable metric overall is the better measure in scenarios (A) and (B) since it delivers optimal results in both.

For walking, when the IMU is placed at the pocket of the upper leg, best results are obtained from any of both independent quaternion metrics that find the matching joint within 83 to 108 samples (or 2.8 to 3.6 seconds). While the standard quaternion distance metrics may have their benefits in calibrated scenes, the calibration independent metrics have their big advantage in environments with unknown setups. Both the stable and normal distance measures have shown to have their specific area of application, depending on the stability of the joint orientation estimates.

3.5 SUMMARY

In this chapter, it was shown how complementary motion sensing from optical and inertial MoCap techniques can be achieved and what considerations have to be taken into account. In a short case study with 10 participants, it was shown how a combined use of both modalities can improve the capturing of a person's wrist and upper arm by mitigating detrimental effects of occlusion events on optical motion data with occlusion-free inertial data. Furthermore, a method was proposed that enables the identification of the person and the limb an inertial sensing device is worn on within a stream of motion data obtained from an observing depth camera. Results of the evaluation show that the correct person and limb can be found within 2.5 to 4.3 seconds, depending on the scenario and the used metric for matching. The proposed method can be considered an important algorithmic component for combining inertial and optical motion data that enables a variety of applications beyond mere motion capturing: The identification makes it possible to establish a communication between the different devices, which then can transmit all kinds of person related data. This ranges from contextual or environmental data of a person's surroundings over a person's position, for instance to enable indoor localization on a wearable device, up to externally measured physiological data, for instance a person's breathing (also see Section 5). Moreover, the literature so far provides only few works that did something coarsely related to identifying a person and limb in a video stream, i.e. few works try to track moving objects or people by correlating inertial data to visual motion in video streams [62, 72]. The proposed method thus closes a gap in this research.

4

COMPRESSING MOTION DATA WITH PIECEWISE LINEAR APPROXIMATION

This chapter is based on the peer reviewed publication [54], where I am the second author of the publication. It has been edited to primarily reflect my contributions to this joint work. The idea of applying piecewise linear approximation on quaternion-based motion data originates from me, while the PLA method fastSW was contributed by Florian Grützmacher (F.G.). F.G. already described fastSW itself in his dissertation [51], but not in the context of compressing quaternion-based motion data. The different authors' contributions, as also stated on the publication, are as follows (with J.K. being me): Conceptualization of fastSW: F.G.; analysis w.r.t. quaternion-based orientation sensor signals: F.G. and J.K.; dataset preparation: J.K.; experimental evaluation: F.G. and J.K.; software: F.G.; formal and experimental analysis of execution time: F.G.; visualization: J.K.; writing: F.G., J.K., K.V.L. and C.H.; supervision: K.V.L. and C.H.

4.1 INTRODUCTION

Many applications require motion trajectories with a high resolution, for instance to avoid visible motion artifacts and to pick up even fine nuances of user motion during a Motion Capturing (MoCap) session, but also to ensure not to miss important signal features in applications such as activity recognition or when combining different modalities to achieve complementary motion sensing. This demand on data requires a high sampling rate on all connected devices and likewise requires them to transmit a substantial amount of motion data. Some of these devices or the system itself, however, may only have limited resources in terms of battery capacity, bandwidth, processing power, or memory. This especially is true for inertial MoCap systems, which often consist of several stand-alone sensor units with wireless communication capacities. For some applications, such as activity recognition, also common commercial wearable devices with an integrated Inertial Measurement Unit (IMU), for instance smartwatches, fitness tracker, or earbuds, are used to track a user's motion and a farseeing usage of system resources is desired. High sampling and communication rates of wireless low-power networks, such as Bluetooth Low Energy (BLE), on the other hand still impose a drastic increase in total device energy consumption [45, 53]. The main drawback in this case is a limited battery capacity. A convenient way to increase the system lifetime at high sampling rates is to reduce the amount of data to be transmitted. Incidentally, a data reduction will also decrease memory requirements. For this reason, a feasible compression scheme is required that efficiently works in environments with limited resources and thus can directly be deployed on the sensing device itself. In this section, such an algorithm will be proposed. It exploits the circumstance that human motion usually is not performed on all limbs simultaneously and that it is rarely performed at high rates of change. Most limbs most of the time will not move a lot and if, these limbs are expected to move at a relatively steady

pace, i.e. with no sudden interruptions or direction changes. Consequently, a considerable amount of energy can be saved if the transmission of sensor samples with little to no additional information is avoided, for instance, if the observed person is standing still or only moves a single limb. The proposed compression algorithm is based on Piecewise Linear Approximation (PLA), a technique that approximates time series signals with linear segments that are guaranteed to be bound by a user-defined upper segment error. PLA techniques have previously been applied successfully to one-dimensional or multi-dimensional data, including IMU data such as acceleration or angular velocity, where it for instance was used to extract and represent the characteristic signal information for the purpose of activity and gesture recognition. Prominent solutions include recognition techniques by dense motif discovery [17], or continuous string and sequence matching algorithms [155, 161], respectively. To date, and to the best of our knowledge, PLA techniques, however, have not been applied to orientation data or to quaternion-based signals in particular in the literature.

A promising scenario of applying piecewise linear approximation on unit quaternions would be the reduction of MoCap data, not only to reduce the amount of data to be transmitted, as discussed above, but also to reduce the file size of MoCap recordings themselves. MoCap files typically comprise an orientation sample for each captured joint and for every single frame, quickly summing up to a significant amount of data on high frame rates. In contrast to that, man-made animations typically only comprise a couple of keyframes that mark a change in limb movement and that are used to interpolate the missing data in between to achieve a fluid animation, often with the help of animation curves. Although consuming more memory than an animation file, a MoCap file thus does not necessarily comprise more information or more detailed motion features. In this scenario, PLA can extract reasonable keyframes from the original MoCap data by rejecting data with little or no information, e.g. from static joints or linear limb movements that can easily be interpolated from key poses. When the data can be compressed directly on the sensing device, furthermore, the required data traffic can substantially be reduced. This not only applies to MoCap scenarios, but also to applications like activity recognition or even telemedicine and rehabilitation, for instance for IMU-based gait analysis such as presented in [22] or [136].

As will be investigated in Section 4.2, unit quaternions require some special attention when being compressed with PLA, such as producing segment points that are a subset of the original sensor samples or producing connected linear segments to enable a successful interpolation of linear segments for the reconstruction of the original signal. Furthermore, a fast and scalable PLA algorithm is required that can run efficiently on wireless sensing devices. While such algorithms exist, these do not adhere to the aforementioned requirements on unit quaternions and, in turn, other existing PLA algorithms that do adhere to these requirements do not provide for an efficient processing of the sensor signals.

To fill this gap in the state-of-the-art, in this chapter, a new online PLA algorithm is proposed that combines efficient and scalable performance with the ability to approximate quaternion-based orientation sensor signals. The novel method will be referred to as fastSW.

4.2 PIECEWISE LINEAR APPROXIMATION OF UNIT QUATERNIONS

When using quaternions to represent 3D orientations, it is of high importance that these quaternions are unit quaternions, i.e. they have a length of one. Otherwise, they have to be normalized before any operation to avoid undesired effects, such as a vector being scaled when rotating it with a non-unit quaternion or yielding incorrect results from spherical linear interpolation. Also, during quaternion multiplications, even small errors and deviations from unit length can accumulate rapidly. Due to floating point precision, a quaternion should also be normalized from time to time when a lot of quaternion multiplications are performed.

This constraint on quaternions plays an important role when applying a PLA algorithm on quaternion-based IMU sensor signals since some algorithms produce segment points that deviate from the original data. This is generally the case with PLA algorithms based on linear regression, such as Connected Piecewise Linear Regression (CPLR) and Swing Filter (SF). They extrapolate segment points from regression lines and thus do not preserve original signal values (also see column POS in Table 2.1). Furthermore, this extrapolation scales the signal unequally among its axes and after normalization, the resulting unit quaternion represents a different rotation. At higher compression ratios, this results in even higher angular deviations as compared to the original data, which understandably is an undesired behaviour. Other PLA methods produce segment points that are a subset of the original data and thus do not introduce such deviations to the sensor signal. One example hereby is the Sliding Window (SW) algorithm. Figure 4.1 depicts the segment points produced by CPLR, SF, SW, and the proposed method fastSW at a similar compression ratio in order to illustrate the differences of these methods. While segment points of CPLR and SF do not necessarily lie on the original data, segment points approximated by SW (or fastSW) are guaranteed to lie on it, as will be explained in more detail in Section 4.3.

SW is an ideal candidate for the compression of quaternion-based motion data. It preserves the original sensor samples in the produced segment points and at the same time is one of the most efficient state-of-the-art PLA algorithms. Another important aspect is that SW produces connected segments, a necessary requirement to ensure a smooth reconstruction of the motion data without any visible motion artifacts between the single segments. The major drawback of SW is its linear execution time and memory complexity per sensor sample with respect to the produced segment lengths (see Table 2.1). This limits its effective compression ratio. A higher compression ratio requires higher segment lengths and thus more memory and processing capacity, which might both be limited, especially on stand-alone sensor devices.

To overcome this limitation, a novel PLA algorithm is introduced in the next section. It is based on SW and leads to mathematically identical PLA results, but at a $O(1)$ complexity in terms of execution time and memory requirements when processing a new sample. Hence, It is referred to as fastSW.

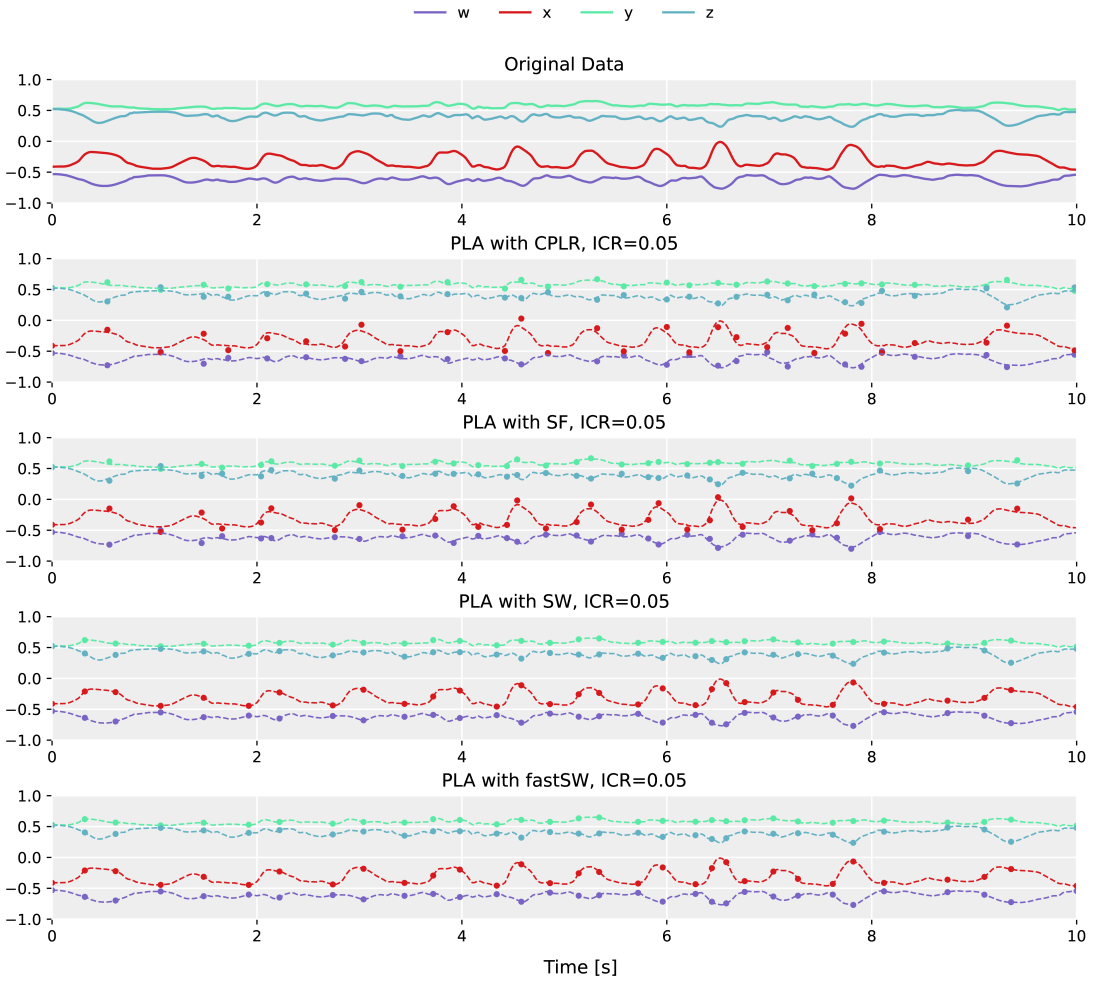

Figure 4.1: Example of the produced segment points of the different PLA methods at a compression ratio of approximately 5% of the original size. Shown are from top to bottom: The original data and, with the produced segment points overlaid, CPLR, SF, SW, and fastSW (the proposed method). The original quaternion data originates from the left shank of the fast-paced *running on spot* activity of user *MR* from the TNT15 dataset [112]. While SW and fastSW create segment points that are a subset of the original data, the segment points of CPLR and SF do not necessarily lie on the original data and moreover deviate on each axis with a different offset [54].

4.3 EFFICIENT PIECEWISE LINEAR APPROXIMATION WITH FASTSW

Attribution: The proposed method fastSW, as described in this section, was conceptualized and implemented by Florian Grützmacher from the University of Rostock and was published in our joint publication [54]. It is described here in detail by myself, using the original formulae and algorithm, for the sake of completeness.

This section introduces a novel PLA method, which is based on the SW algorithm and in the following will be referred to as fastSW. Analogous to SW, it creates for each new (n-th) sensor sample within a series of measurements a temporary segment that approximates the original signal with a linear function $\beta \cdot t$. The variable β hereby represents the slope vector of the D-dimensional segment and t describes the length

in time of that segment, respectively. Each slope vector entry β_d can be computed from the value of the current sensor sample s_n, the value of the segmentation point $\tilde{s}_{i-1}$ of the last segment $i-1$, and the time $t = \tau(s_n) - \tau(\tilde{s}_{i-1})$ between them, as defined in (4.1):

$$\beta_d = \frac{v(s_n, d) - v(\tilde{s}_{i-1}, d)}{t} \quad , \qquad d = 1, \ldots, D \tag{4.1}$$

The index n of the most recent sample within the currently developing segment is reset to one each time a new segment point i is created since this sample will be set as the first sample of the next segment $i+1$.

To know when a new segment with a different slope needs to be created, a measure for the error between the approximating segment and the original data is required. In the SW algorithm, this error is based on the Sum of Squared Residuals (SSR) between the current segment $(\tilde{s}_{i-1}, s_n)$ and the respective part of the original signal. It is defined as:

$$SSR_n = \sum_{j=1}^{n} \sum_{d=1}^{D} \left(y_{d,j} - \beta_d \cdot t_j\right)^2 \tag{4.2}$$

The variables $y_{d,j}$ and t_j are relative to the current (i-th) segment's start and represent the amplitude and the timestamp of the j-th sensor sample within the current segment. Both are computed as $y_{d,j} = v(s_j, d) - v(\tilde{s}_{i-1}, d)$ and $t_j = \tau(s_j) - \tau(\tilde{s}_{i-1})$.

To compute the segment's SSR, the SW algorithm buffers all original samples and iterates over these each time a new sample is added, which leads to a linear time complexity at this point. To achieve a constant time update of the segment's SSR error, fastSW follows a different approach. Equation (4.2) can be reordered by applying binomial expansion and swapping the two commutative sums to yield Equation (4.3):

$$SSR_n = \sum_{d=1}^{D} \left(\sum_{j=1}^{n} \left(y_{d,j}^2\right) - 2\beta_d \sum_{j=1}^{n} \left(t_j \cdot y_{d,j}\right) + \beta_d^2 \sum_{j=1}^{n} \left(t_j^2\right) \right) \tag{4.3}$$

The outer sum over the signal's dimension D is independent of the segment length and the individual inner sums over n now can be updated in constant time. To avoid numerical issues on large sums that span a wide range of values, the inner sums of Equation (4.3) furthermore are substituted by corresponding mean values. These mean values are updated instead and yield the actual value of the respective sum when multiplied by n. They are defined as:

$$\sum_{j=1}^{n} \left(y_{d,j}^2\right) = \overline{y^2}_{d,n} \cdot n \tag{4.4}$$

$$\sum_{j=1}^{n} \left(t_j \cdot y_{d,j}\right) = \overline{ty}_{d,n} \cdot n \tag{4.5}$$

$$\sum_{j=1}^{n} t_j^2 = \overline{t^2}_n \cdot n \tag{4.6}$$

The means $\overline{y^2}_{d,n}$, $\overline{ty}_{d,n}$, and $\overline{t^2}_n$ themselves are updated as:

$$\overline{y^2}_{d,n} = \overline{y^2}_{dn-1} + \frac{\overline{y^2}_{dn-1} - y^2_{d,n}}{n} \tag{4.7}$$

$$\overline{ty}_{d,n} = \overline{ty}_{dn-1} + \frac{\overline{ty}_{dn-1} - t_n \cdot y_{d,n}}{n} \tag{4.8}$$

$$\overline{t^2}_n = \overline{t^2}_{n-1} + \frac{\overline{t^2}_{n-1} - t_n^2}{n} \tag{4.9}$$

By substituting the sums by the means, Equation (4.3) now becomes:

$$SSR_n = \sum_{d=1}^{D} \left(\overline{y}^2_{d,n} - 2\beta_d \overline{ty}_{d,n} + \beta_d^2 \overline{t^2}_n \right) \cdot n \tag{4.10}$$

Equation (4.10) allows an update of the segment error SSR_n that only has a linear time complexity with respect to the signal's dimension D. Since Quaternion-based sensor signals have a constant dimension of $D = 4$, the update of the segment error also becomes constant. This difference in the computation of the SSR distinguishes fastSW from SW. The fastSW algorithm thus can be implemented by substituting the SSR computation of the SW algorithm with the newly developed approach using Equations (4.7) to (4.10). The pseudo-code of fastSW can be found in Algorithm 1.

Algorithm 1 fastSW [54].

```
1:  procedure PROCESS_SAMPLE(sample s, segment array S̃[], index i)
2:      n := n + 1
3:      SSR_n := 0
4:      t_n := timestamp(s) − timestamp(S̃[i − 1])
5:      for d in (1, ..., D) do
6:          y_n[d] := value(s, d) − value(S̃[i − 1], d)
7:          β[d] := y_n[d]/t_n[d]
8:          SSR_n := SSR_n + (ȳ²_{n−1}[d] − 2β[d] · t̄y_{n−1}[d] + β[d]² · t̄²_{n−1}) · (n − 1)
9:      if SSR_n <= TH then
10:         t̄²_{n−1} := t̄²_{n−1} + ((t_n · t_n) − t̄²_{n−1})/n
11:         for d in (1, ..., D) do
12:             t̄y_{n−1}[d] := t̄y_{n−1}[d] + ((t_n · y_n[d]) − t̄y_{n−1}[d])/n
13:             ȳ²_{n−1}[d] := ȳ²_{n−1}[d] + ((y_n[d] · y_n[d]) − ȳ²_{n−1}[d])/n
14:         s_{n−1} := s
15:         return 0
16:     S̃[i] := s_{n−1}
17:     s_{n−1} := s
18:     n := 1
19:     t_n := timestamp(s) − timestamp(S̃[i])
20:     t̄²_{n−1} := t_n · t_n
21:     for d in (1, ..., D) do
22:         y_n[d] := value(s, d) − value(S̃[i], d)
23:         t̄y_{n−1}[d] := t_n · y_n[d]
24:         ȳ²_{n−1}[d] := y_n[d] · y_n[d]
25:     return 1
```

The array of segment points $\tilde{S}[]$ in Algorithm 1 needs to be initialized with the very first sensor sample since the first sensor sample will also be the first segment point in the list. Consequently, the function `PROCESS_SAMPLE` is not invocated for the first sample, but each time a new sample s_j arrives. The function's parameters are the current sensor sample s_j to be processed, the segment array $\tilde{S}[]$, and the index i, referencing the end of the segment array where the next segment point will be stored on success.

The variables n, TH, D, $\overline{ty}_{n-1}[]$, $\overline{y^2}_{n-1}[]$, and s_{n-1} are global, while all other variables are temporary. The user defined variables TH and D represent the error threshold at which a new segment will be started and the dimensionality of the sensor signal, respectively, and need to be set by the user accordingly. A higher threshold will likely lead to a higher compression. All other global variables need to be initialized with zero. The function returns zero to indicate that no new segment point was created. Otherwise it returns one. In the latter case, the index i needs to be incremented by one before the next invocation of `PROCESS_SAMPLE`.

4.4 EVALUATION

For the evaluation, a selection of the most efficient state-of-the-art methods that provide PLAs with connected segments are implemented to process 4-dimensional quaternion data (see Table 2.1). Namely, these are CPLR, SF, SW, and the proposed method fastSW. Although CPLR, SF, and fastSW do not require a segment length limitation, all methods are implemented to have a maximum segment length of 1000 samples, which is enforced to ensure comparable results to SW. Consequently, the minimum achievable compressed data size will be 0.1% of the original data size.

The evaluation consists of two parts: In a first experiment, the approximation quality of all methods is assessed and compared on a public dataset. The same dataset then is used in a second experiment to investigate the computational complexities of the different PLA algorithms in a realistic scenario, including a Worst-Case Execution Time (WCET) analysis on a representative architecture for wearable devices.

4.4.1 *Dataset and Experiment Design*

The proposed method is experimentally evaluated on the publicly available TNT15 dataset [112]. It comprises 4 actors performing 7 different activities, summing up to a total of 28 distinct recordings with 4040 to 10,180 samples per file. The activities include walking, running on the spot, rotating arms, jumping and skiing exercises, dynamic punching, and two not further specified, random activities. Each recording was performed with 10 IMUs that are attached to the shanks, thighs, lower arms, upper arms, neck, and hip. Each sensor provides acceleration and quaternion-based orientation data at a sampling rate of 50 Hz. For the evaluation, the acceleration data is discarded.

The different PLA algorithms can only be compared on equal or at least similar compression ratios, which the different PLA methods are not guaranteed to reach when processing the same dataset. For this reason, each algorithm has to be executed multiple times on the entire dataset, each time with a different threshold value to ensure an appropriate range of achieved compression ratios at which they can be compared. This is achieved by approximating each of the 28 recordings in the TNT15 dataset at 205 different threshold values, starting at a value of 0.000001 and logarithmically increasing evenly among 9 magnitudes up to a value of 1000.

4.4.2 *Visual Inspection*

Figure 4.2 illustrates the reconstruction results of the different PLA algorithms at five different compression ratios on the *running on spot* activity of user *MR* from the TNT15 dataset. For each method and compression ratio, it shows the same six consecutive frames. They are obtained by using Spherical Linear Interpolation (SLERP) to interpolate between the respective segment points with a time step of 100 ms between the single frames.

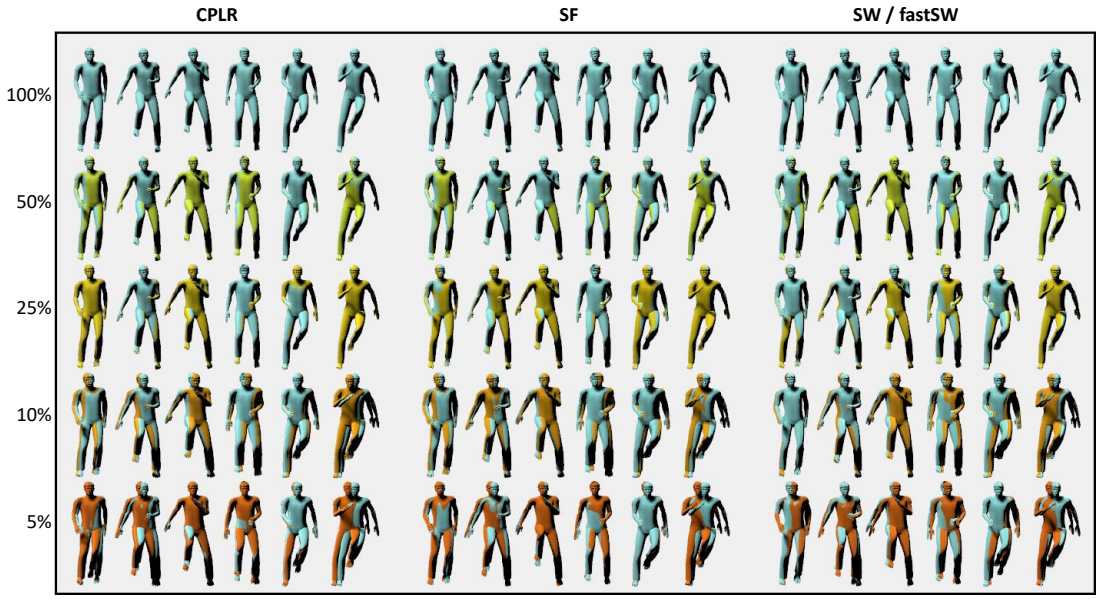

Figure 4.2: Visualization of the reconstruction results of the fast-paced *running on spot* activity of user *MR* from the TNT15 dataset using the methods (from left to right): CPLR, SF, and SW/fastSW. Single frames (from left to right) are taken at six different, successive time points every 100 ms. Colors indicate the compression ratio and are arranged from top to bottom, with: The original file (light blue), 50% (light green), 25% (yellow), 10% (orange), and 5% (dark orange) of the original file size. Ground truth frames (light blue) are overlaid over all frames to highlight deviations. At 50% and 25%, a good approximation is obtained from all methods, while higher compression ratios yield less accurate reconstructions, independent of the method used.

Figure 4.2 is intended to provide an insight in the visual quality of the reconstructed motion data at different compression ratios and with different PLA algorithms. The general trend is that the reconstructed frames deviate more from the ground truth frames as the compression increases. This trend is independent of the method used and illustrates the trade-off between approximation error and compression ratio. Although SW and fastSW, in contrast to CPLR and SF, guarantee that their segment points are a subset of the original data, large differences between CPLR, SF, and SW or fastSW are not directly visible from the visual inspection alone. The differences between the methods will further be evaluated by looking at the data itself in the next section.

4.4.3 *Approximation Quality*

To estimate the overall approximation quality of the different PLA algorithms, the Average Angular Deviation (AAD) between the reconstructed signal and the original signal is used. It is computed by summing up the geodesic angles between the original sensor samples and their respective approximations and, to account for different signal lengths, by dividing the result by the number of samples m in the original sig-

nal. The geodesic angle between two quaternions q_0 and q_1 can be computed with the quaternion dot product $\langle q_0, q_1 \rangle$ as:

$$\Delta(q_0, q_1) = 2 \arccos |\langle q_0, q_1 \rangle| \tag{4.11}$$

The AAD then is defined as:

$$AAD = \frac{1}{m} \sum_{i=0}^{m-1} \Delta(S[i], \tilde{S}'[i]) \tag{4.12}$$

$\tilde{S}'$ hereby is the reconstructed signal obtained by interpolating between the segment points of the compressed PLA signal $\tilde{S}$ at the corresponding timestamps of the original signal S using SLERP. SLERP hereby has a few benefits over other interpolation techniques: It is a standard method for interpolating quaternions, retains their unit length, gives a reasonable approximation because it interpolates along the geodesic of the quaternion hypersphere, i.e. the shortest angle between two orientations, and it conserves the angular velocity of the resulting rotation.

Of interest for the evaluation is the approximation quality that can be achieved at a certain compression ratio. The Inverse Compression Ratio (ICR) is defined as the division of the length of the compressed signal $\tilde{m}$ by the length of the original signal m, as stated in:

$$ICR = \frac{\tilde{m}}{m} \tag{4.13}$$

A lower ICR means that fewer segment points are created and thus a higher compression could be achieved. In general, it can be assumed that a lower ICRs leads to a higher approximation error, because there is a trade-off between both.

The approximation error can be plotted over the ICR, with the benefit that the distance of the resulting curve to the origin directly indicates the approximation quality of the respective algorithm. The closer it is, the better is the respective approximation quality. The use of a wide range of threshold values for the approximations ensures a sufficient amount and a good coverage of the ICR ranges, effectively enabling a good comparison between the algorithms.

For this reason, the TNT15 dataset has been approximated with CPLR, SF, SW, and fastSW using a wide range of threshold values (see Section 4.4.1). As result, for each algorithm, a wide range of angular deviations with the respectively achieved ICR is obtained. From these values, the long term average, standard deviation, and maximum of the approximation error are computed and plotted over the respective ICR as depicted in Figure 4.3. It shows how well the different PLA algorithms operate on a variety of different movements performed by real humans wearing inertial sensing devices.

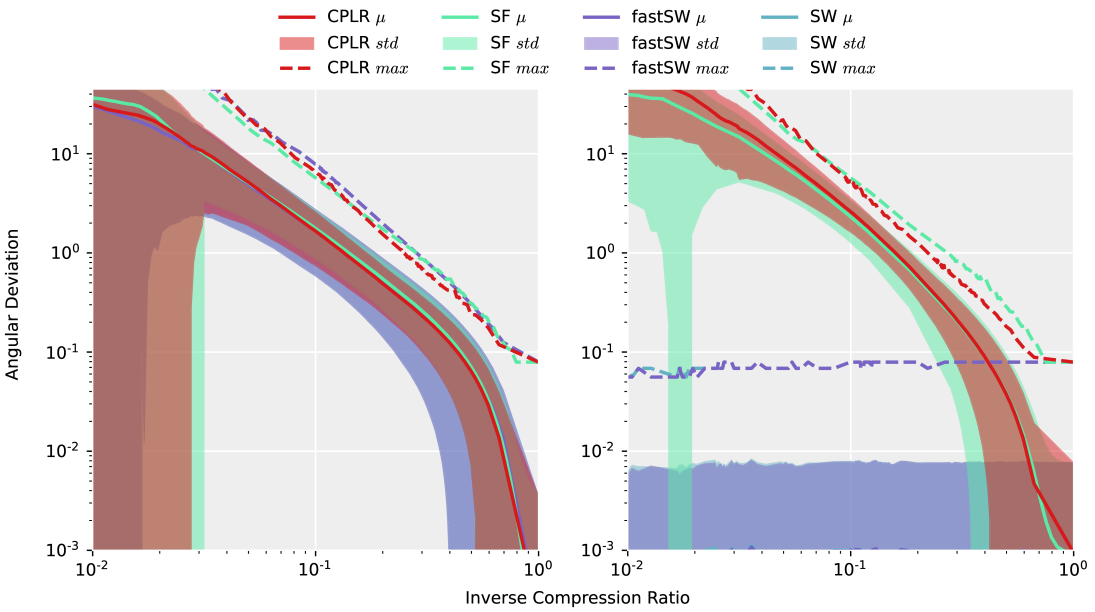

Figure 4.3: Average (solid line), standard deviation (shaded area), and maximum (dashed line) of angular deviations plotted against the respective ICR as obtained by CPLR, SF, SW, and fastSW after compressing the entire TNT15 dataset with a high range of different threshold values to visualize the respective approximation quality of the different methods. Colors of the shaded regions (standard deviations) mix up and, due to the logarithmic scale of the y-axis, extend over the whole plot at smaller ICRs. **Left**: Approximation quality of the reconstructed signal as compared to the whole original signal after interpolating the produced segment points using spherical linear interpolation. **Right**: Approximation quality as compared to the original signal, but only at the segment points created by the different algorithms. While differences of the methods on the reconstructed signal (left plot) are not easy to spot, they are obvious when looking at the segment points themselves (right plot). SW and fastSW show over the whole range of ICRs only small, almost constant approximation errors, with slight angular deviations caused by numerical precision issues. The approximation errors of the segment points of CPLR and SF on the other hand are heavily influenced by the compression ratio. A decreasing ICR causes increasing angular deviations [54].

Figure 4.3 (left) shows the angular deviations of all approximated data points after interpolation using SLERP, while Figure 4.3 (right) only shows the angular deviations of the segment points themselves, without interpolating between them. As expected, SW and fastSW exhibit the same or nearly the same approximation quality on both plots due to the same underlying concept to generate PLA segments. Small deviations are caused by floating point precision in combination with a different mathematical approach to compute the segment points.

In Figure 4.3 (left), the plot of interpolated data points, all methods feature a similar approximation quality and differences of the methods are not obvious. This changes when looking at the segment points only. In Figure 4.3 (right), a significant difference is visible. While SW and fastSW only show small angular deviations that, independent of the compression ratio, on average stay below 0.001°, CPLR and SF show a dependency where with decreasing ICR the angular deviations of their segment points increase. On an ICR of 0.01, the average angular deviation of their

segment points reaches up to more than 40°. The main difference between the methods is, that SW and fastSW create segment points that, beside numerical precision, do not deviate from the original data, while CPLR and SF, both linear regression-based algorithms, create extrapolated segment points that do deviate, especially at lower ICRs, i.e. when the compression increases. The angular deviations of CPLR and SF are caused by an unequal scaling of the quaternion's axes, effectively not only changing its unit length, but also rotating it.

In summary, the correct segment points of SW or fastSW have less impact on the approximation quality than anticipated, because they cannot compensate for missing data and signal features in between them. Also, just the high amount of interpolated data necessarily created on higher compression ratios causes a higher impact on the approximation error than incorrect, but favorably placed segment points that are used to interpolate the missing data. Accurate segment points on the other hand still are preferable, especially because they do not need to be normalized and do not introduce additional rotation errors, e.g. when working on the segment points alone.

4.4.4 *Execution Time Analysis*

Attribution: The execution time analysis, as described in this section, was performed by Florian Grützmacher from the University of Rostock and was published in our joint publication [54]. It is described here to illustrate the benefits of the method.

To assess the different PLA methods' computational performance on real motion data, the execution time for processing the entire TNT15 dataset is measured. The measurements are based on taking timestamps before and after the execution of the respective PLA function to process a single sample by using the `clock_gettime` method with the `CLOCK_MONOTONIC_RAW` as clock source and a 1 ns time resolution. Since in this case only the average execution time to process a sample with respect to the average segment length is relevant to assess the computational complexity, the experiments are performed on a standard `x86_64` architecture with a Linux operating system in kernel version `5.12.9` and an Intel Core i7-5600U processor. The algorithms, furthermore, are compiled using the GNU Compiler Collection (GCC) C compiler with version 11.1.0 [157].

Figure 4.4 depicts the average execution time over the average segment length of the different PLA algorithms CPLR, SF, SW, and fastSW after processing the entire TNT15 dataset on a wide range of different threshold values (see Section 4.4.1). Note that higher average segment lengths are desired because they yield a higher compression. The average execution time of CPLR, SF, and fastSW clearly is independent of the segment length and, apart from some outliers at small segment lengths, is constant over the whole range of segment lengths. The outliers mostly are caused by cache misses as well as other architectural influences. SW on the other hand shows with increasing segment lengths a linear growth in execution time (note the logarithmic scale of the time axis). In direct comparison, fastSW furthermore exhibits the lowest execution time of all methods, apart from segment lengths of around 2 or less samples, where SW is faster.

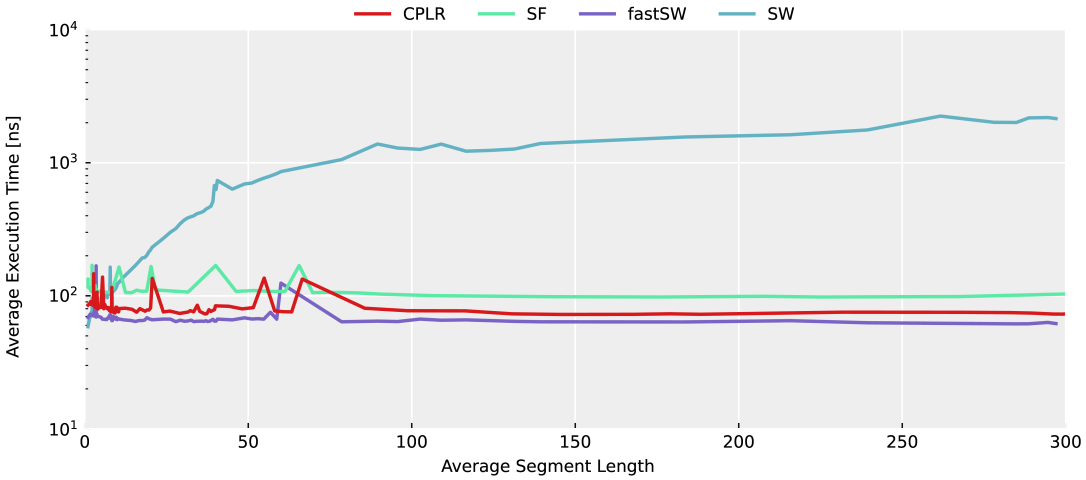

Figure 4.4: Average execution time per sample over the resulting average segment length on a `x86_64` architecture. Values are obtained from approximating the entire TNT15 dataset with the respective PLA algorithm using a wide range of different threshold values. The execution time is plotted on a logarithmic scale. SW shows a linear dependency of the average execution time with respect to the average segment length, while CPLR, SF, and fastSW exhibit a constant execution time over the whole range of segment lengths, with a few outliers caused by cache misses and other external influences [54].

To validate the computational complexity of CPLR, SF, SW, and fastSW on wearable devices, a WCET analysis is performed on the ARM Cortex-M4 architecture, which is representative for wearable devices. Each algorithm therefore has been compiled for the Cortex-M4 target platform using the Arm Embedded GCC (version 11.1.0) [157], with floating-point unit specific calling conventions and the highest optimization level with respect to to execution time. Here, the instruction count is used as a performance measure, because it is independent of the data and, more importantly, of the clock frequency of the specific microcontroller. To this end, the instruction counts of the different methods are manually assessed from the compiled code and are summarized in Table 4.1.

Table 4.1: Instruction counts (IC) of CPLR, SF, SW, and fastSW on an ARM Cortex-M4 microcontroller. Implemented to process 4-dimensional quaternion data. n denotes the current segment length [54].

Algorithm	**min. IC**	**max. IC**
CPLR	198	209
SF	252	420
SW	53	$161 + n \cdot 35$
fastSW	191	210

In contrast to SW, which has 35 additional instructions for each of the n samples currently in the buffer, the instruction counts of CPLR, SF, and fastSW are independent

of the segment length. SF furthermore has in general a higher instruction count than CPLR or fastSW, with up to twice as much in the worst case scenario.

4.5 DISCUSSION

In terms of approximation quality, a clear advantage of fastSW over other state-of-the-art PLA algorithms cannot be stated. The shear amount of more or less erroneously approximated data on high compression ratios simply outweighs the few correctly placed segmentation points. However, there are scenarios where a reduction of orientation sensor data is desired and, at the same time, accurate supporting points are necessary. This is especially the case when working on the segmentation points alone, e.g. when using them as feature points that mark a change in user movement such as in activity recognition applications. Moreover, having accurate segment points that do not have to be normalized and that do not introduce additional rotation errors also is beneficial when combining optical and inertial motion capturing.

In direct comparison to SW, fastSW does not necessarily provide the same segmentation points due to small differences in the segment error calculation caused by a different order of otherwise identical mathematical operations in the presence of limited floating point precision. The difference, however, is minimal.

The TNT15 dataset comprises many fast paced movements where many limbs are moved simultaneously and does not necessarily reflect real world scenarios with calmer periods in between. A higher compression ratio while achieving the a similar approximation quality thus likely is achievable in other scenarios. The opposite, however, can also be the case, e.g. when monitoring fast paced sport activities.

4.6 SUMMARY

So far, there was no computationally efficient PLA algorithm that is suitable for the compression of unit quaternions. An analysis revealed that such an algorithm is required to produce connected segments and that the produced segment points are a subset of the original data. Methods that violate the latter constraint have been shown to be less suitable, because (1) the resulting quaternions on the segment points need to be normalized to restore their unit length, and (2) even slight deviations from the original data on different axes can cause significant angular deviations. Methods that do create segment points that are a subset of the original data, such as SW, on the other hand, are not as efficient in terms of time and memory complexity. This is where fastSW jumps in. It closes a gap in the state-of-the-art by being suitable for the compression of unit quaternions while achieving a time and memory complexity of $O(1)$ with respect to the compression ratio. This efficiency makes fastSW well suited to be deployed on embedded systems with limited resources, such as stand-alone inertial sensing devices.

In light of the complementary motion sensing approach as described in Chapter 3, fastSW thus fits well as a compression scheme that allows an efficient communication of motion data between the various sensing systems that are involved in the process, be it from a body-worn inertial sensing device to an external camera system or vice versa.

5 REMOTE RESPIRATION ESTIMATION

This chapter is based on the peer reviewed publications [79], [80], and [81]. Some passages have been quoted verbatim. I am the first author of all three publications. Section 5.6.2 is based on the peer reviewed publication [21]. The idea to use the system in the context of e-health and telemedicine originates from me and I extracted the core elements of this idea from the publication. The study experiments were conducted by Steffen Brinkmann and the full evaluations and details can be found in the original work.

5.1 INTRODUCTION

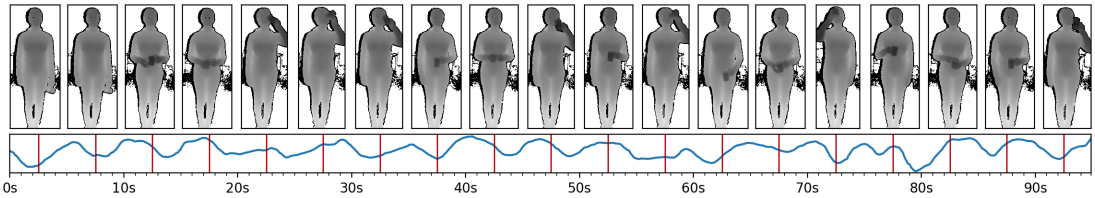

Figure 5.1: An example of depth-based respiration estimation, taken from a distance of 2 meters. The different frames of a depth recording are drawn on the top row, starting at 2.5 seconds with an equal spacing of 5 seconds. The breathing signal as estimated from a model-based approach is drawn on the bottom row. Red markers indicate the time points at which the respective frame above the marker was captured. In general, the user can be anywhere in the frame. The proposed approach allows the user to perform activities that self-occlude the torso. An overview of the model-based as well as various other types of state-of-the-art depth-based respiration estimation methods will be presented in Section 5.5.1.

Respiration is the physiological process of our body to exchange carbon dioxide with oxygen. Inhalation mainly happens through actively contracting the diaphragm and increasing the thoracic cavity, while exhalation typically occurs as a passive process due to the elasticity of the lungs. In contrast to most other vital body functions, respiration can be controlled consciously. Unconscious breathing on the other hand is controlled by the respiratory centers of the brainstem that regulate the respiratory rate mainly depending on the pH of the blood. According to [100], three modalities can be used to measure human respiratory rate: Measurements can be based on other physiological signals, on respiratory movements, or on airflow. Monitoring a subject's respiration plays an important role in medical diagnosis and treatment [5, 34] as it tends to not only change with physical exercise, but also with a range of conditions like fever and illness [129]. Although the human respiratory rate is an important vital parameter, it is still under-measured [41, 153]. Typical medical applications, to name a few examples, are sleep monitoring or asthma therapy. In asthma therapy, for instance, patients usually have to go to a special lab or their doctor's office to have their respiration monitored. Since in these places only a limited amount of time and

space is available [19], patients could benefit from long-term observations of their breathing made at home, given a suitable sensing device is available to them. Respiration furthermore is closely linked to behavioural and affective states [118] and can serve as an indicator of wakefulness or concentration due to the circumstance that the breathing rate decreases when drifting towards sleep [58]. Beyond medical applications like sleep assessment or asthma therapy, in sports and fitness applications as well as in well-being, mindfulness, and meditation exercises the respiratory rate often is used to assess a subject's performance or is used to induce or control a specific state of the body or the mind. Likewise, in these scenarios the user often is required to maintain a specific breathing pattern and needs to rely on external feedback that might be improved given a suitable sensing device.

Conventional sensors like mask-like spirometers, nasal tubes, respiration belts worn around the chest, or skin-based photoplethysmography, but also more recently proposed methods utilizing body-worn inertial sensors, like [59], require physical contact to the user's body and, over longer time periods, tend to become uncomfortable or restraining for the person to wear. Especially in fitness applications, but also in scenarios where users perform breathing exercises for stress reduction, like for instance meditation, such devices should be easy to set up, comfortable to wear, non-obtrusive, and not cause distraction as these conditions might lower their acceptance. Respiration estimation based on photoplethysmography furthermore can only measure heart rate or oxygen saturation of the blood reliably, while respiration is derived from heart rate variability. Consequently, these methods lack a reliable and instant respiratory rate estimation [134].

Several methods have been proposed to estimate a person's breathing through ambient sensing, eliminating the need of any body-worn devices. Yet, it is notoriously difficult to obtain a respiration signal from a distance. Methods based on a depth camera picking up the tiny changes in distance of the chest or abdomen during respiration hereby have shown promising results [140]. Most such depth-based methods, however, are designed for certain, well-defined scenarios and the assumptions and conditions in which they were evaluated have remained limited and far from realistic: Cameras had a direct line of sight to the user's torso, and users cannot perform activities other than lying down or sitting still. Furthermore, they lack a systematic evaluation of important parameters and conditions, such as distance to the camera, the observed body region, or the user introducing subtle body movements while for instance standing upright. User studies often are conducted with only a few participants and a quantitative comparison to different methods is not available. It therefore remains widely unclear how the existing methods perform under various conditions and how they compete.

To overcome many limitations of current state-of-the-art depth-based respiration estimation methods, a novel approach to monitor a user's respiration from a depth camera is proposed. It does not require the user to lie down or sit still, nor to stay at a predetermined position or distance to the camera, and it is robust against small body movements and occasional occlusions of the user's upper body with its arms. An example of this method in action is presented in Figure 5.1. It is introduced in Section 5.2 and will be validated in a user study in Section 5.4.

Furthermore, in Section 5.5, a detailed and systematic in-depth analysis of the impact of many important parameters and conditions is performed on the proposed, as well as on the most common state-of-the-art depth-based respiration estimation techniques. The parameters under evaluation include the distance between user and the camera, the observed body region, the pose and activity, the user's respiratory rate, its gender, as well as user specific influences. The analysis is followed by a discussion of the circumstances under which any of these methods has the most advantages to be used. As already stated above, the influence of all these parameters on the various state-of-the-art methods to date remains unknown. This study thus aims at closing this gap in the research.

5.2 PROPOSED METHOD

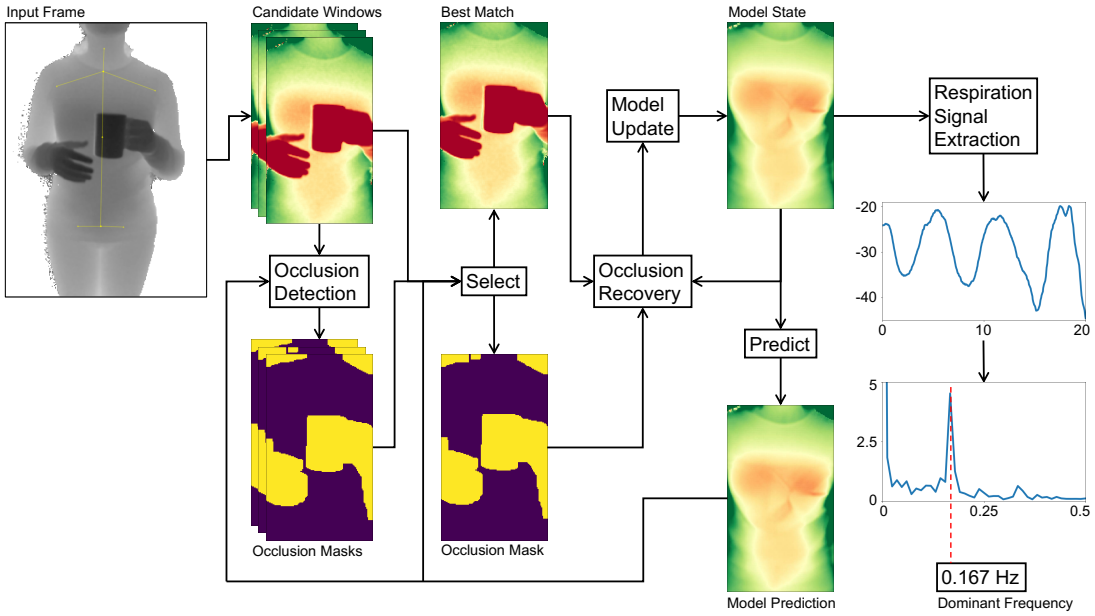

Figure 5.2: The core steps of the proposed respiration monitoring method. From left to right: The process starts with the camera's depth *input frame* and the estimated torso position of the user. Since joint position estimates contain jitter, multiple *torso window candidates* are selected. Combined with the torso prediction from the previous frame, each candidate is then assigned an occlusion mask. The best matching candidate to the torso prediction is selected as *torso window* and forwarded to the occlusion recovery stage, which uses the occlusion mask as well as the torso prediction from the previous frame to yield the current *torso model*. The torso model then delivers the prediction for the next frame, where it will be used for occlusion recovery. The torso model is transformed to a single respiration state value, the history of these values yields a respiration signal that, after Fourier Analysis and extraction of the *dominant frequency*, estimates the respiratory rate.

This section gives an overview of the proposed method to monitor a user's respiration from subsequent depth camera frames. The method focuses on an indoors setting where a user is facing a depth camera, which also tracks the user's body joints. The data processing chain of the method is divided into different stages that

each perform an abstraction from the individual input depth frames to the final respiratory rate estimates. A sketch of the overall process is provided in Figure 5.2.

Starting from the camera's single raw depth frames, users and their postures are first identified. For each user, the body's joint positions are used to mark the user's torso (Section 5.2.1), with especially the respiration-related regions of the chest and abdomen being of high interest. With the help of the model, occluded parts can be identified and masked out, and the alignment of the torso window in case of a mismatch of the estimated joint positions can be refined. The selected region and its occlusion mask (Section 5.2.2) subsequently are forwarded to the occlusion recovery stage. Here, all occluding parts from the depth image are replaced by a model-generated approximation of the torso surface, also removing any salt and pepper noise present in the depth image. The individual steps of the occlusion recovery are presented in Section 5.2.3 and the adaptive model implementation is specified in Section 5.2.4. The resulting recovered depth image satisfies the criteria of being well aligned and keeping sufficient detail in occluded areas such that it can be used to update the model without losing its integrity. The following stage extracts the respiration signal in a movement-robust fashion by identifying the torso deformation caused by breathing as described in Section 5.2.5. From the resulting breathing signal, the respiratory rate can finally be estimated through Fourier analysis.

5.2.1 *Locating Users and Torso Windows*

The incoming depth frames are first scanned for users present in the scene and, if any are detected, the users' torso areas are tracked over time to extract respiration-relevant motions. Figure 5.3 illustrates the process of detecting the torso region using body joint estimates. Potential misalignments due to arm movement hereby pose a challenge. In the following, the overall process of identifying the torso region correctly and in a stable way is introduced.

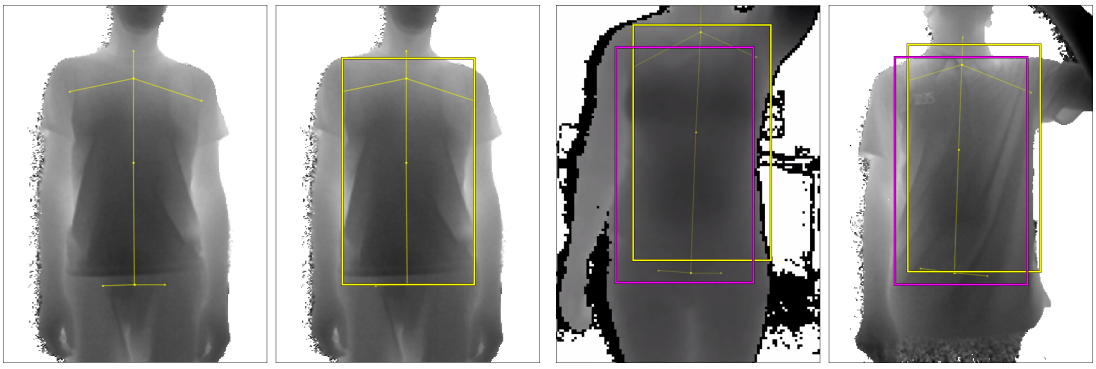

Figure 5.3: The first image to the left illustrate the estimated joint positions of the shoulders, hip, neck, and spine with connections in yellow drawn on top of the input frame. In the second image, the torso window as obtained from the joint positions is drawn as yellow rectangle on top of the joints. Here the window is correctly aligned. The two images to the right illustrate window misalignment due to arm movement, with magenta being the torso window alignment as required by the model, and yellow being the torso window as estimated by the joint positions.

Input Frame. Tracking is done by a skeletal model consisting of 25 joints, revealing the person's position and pose, following the approach of [150]. Here, the implementation in the Kinect SDK 2.0 is used, which is capable of tracking up to six persons simultaneously. Knowing a user's body posture enables the tracking of the *torso window*, which reveals slight breathing motions across the torso surface during respiratory cycles. The decisive anchor points to determine the torso position and scale horizontally are the left and right shoulder joints. In the vertical direction, the neck, the shoulder mid, the spine mid, and hip joints are used. The estimated joint positions can be expected to be slightly unstable over time, jumping between neighbouring pixel positions. In some cases, in the presence of occlusions or arm movement for instance, the joints may also be misaligned (also see Figure 5.3). Especially when users are further away, such jumps or misalignment occurrences are more likely.

Torso Prediction. The depth input frames typically comprise several challenges, including noise, different scales and extents of the user's torso region, background and outlier pixels, occlusions and shadowing, motion artifacts, and surface deformations such as folds caused by clothing. To overcome these challenges, an adaptive model is maintained for each user, which estimates the torso window appearance over time. More details on how this model is constructed and updated will be given in 5.2.4.

Torso Window Candidates. Some heuristics are used to stabilize the tracking of the torso window. First, the window's aspect ratio is fixed to a predefined value at initialization. The torso window's change in size caused by small swaying movements while standing is negligible, so a fixed height and width are assumed across frames. To cater for users moving strongly towards or away from the depth camera, several rescale operations would need to be performed to fit the window in the next step. Second, the position of the torso window, although allowed to move in the frame, is clamped to a position with similar content of the predicted depth image from the model. For this, all possible x_{win} and y_{win} positions that fall in between the current and previous window position are permuted to form as many different *candidate windows* for the torso. All candidates obtained from $I(x_{win}+x, y_{win}+y)$ are compared pixel-wise to the model prediction image $I_{predict}(x,y)$ as defined in (5.1):

$$I_{best\,fit} = \underset{x_{win},y_{win}}{\arg\min} \sum_{x,y\in\mathbb{N}}^{w,h} M_{win}^{-1}(x,y)\cdot(I(x_{win}+x, y_{win}+y) - I_{predict}(x,y))^2 \quad (5.1)$$

By multiplying the energy term with an inverted occlusion mask $M_{win}^{-1}(x,y)$ of the respective candidate window win, all detected occlusion pixels are ignored, as they unnecessarily increase the difference of both images, pushing the window away from the occlusion instead of matching the surface. The best matching window is selected as the torso window and forwarded to the *Occlusion Recovery* step. The next section will first detail the procedure of selecting an occlusion mask.

5.2.2 *Occlusion Mask*

The presence of occlusions poses a particularly hard problem for optical respiration monitoring, as (1) important regions of the torso may be blocked, and (2) movement

of the occluding entity may be misinterpreted as evidence for a respiration signal. It is, therefore, important to detect and mask any entities that occlude the torso with an occlusion mask M. This mask helps maintaining a breathing-relevant set of depth pixels and ensures the integrity of the model.

Since occluding entities will be in front of the torso, their corresponding depth values will always be smaller than those of the torso. So, to identify occluding entities, knowledge about the user's torso surface in the current frame is required. Fortunately, the model predicts the surface appearance and distance of the tracked torso, which, when being subtracted from the corresponding input frame, yields negative values on potential occlusions in the obtained difference image. To find the occlusion mask, each pixel in the difference image then only needs to be compared to a certain threshold that defines the minimum distance to the torso surface. Anything below this threshold is considered an occlusion and anything above is considered part of the user's torso surface, including skin deformations or clothing. The occlusion threshold has to take into account that the model has a small delay due to the model's low-pass filtering behaviour, where fast movement may lead to incorrect masking. The same applies to noise in the input frame and small deformations of the torso surface, e.g. due to clothing. In initial studies, an optimal value of 30 mm was found for this threshold. Objects that do erroneously find their way into the model will be excluded from the model as soon as the blocked torso surface becomes visible again. Equation (5.2) mathematically defines the occlusion mask $M(x, y)$ at pixel positions x and y with the input image $I(x, y)$ and the torso prediction $I_{predict}(x, y)$ as well as the threshold value $z_{threshold}$:

$$M(x, y) = \begin{cases} 1 & \text{if } I(x, y) - I_{predict}(x, y) < z_{threshold} \\ 0 & \text{else} \end{cases} \tag{5.2}$$

Occluding entities in depth imaging often show a halo on their edges, caused by interference from the emitted infrared light being reflected from the object upon the torso surface. To avoid such a halo leading to undesirable effects in the next steps, the mask is enlarged by a margin of several pixels, depending on the distance to the camera and the resolution of the torso frame. With this mask, occluding entities can be removed in the torso window as well as in the torso model.

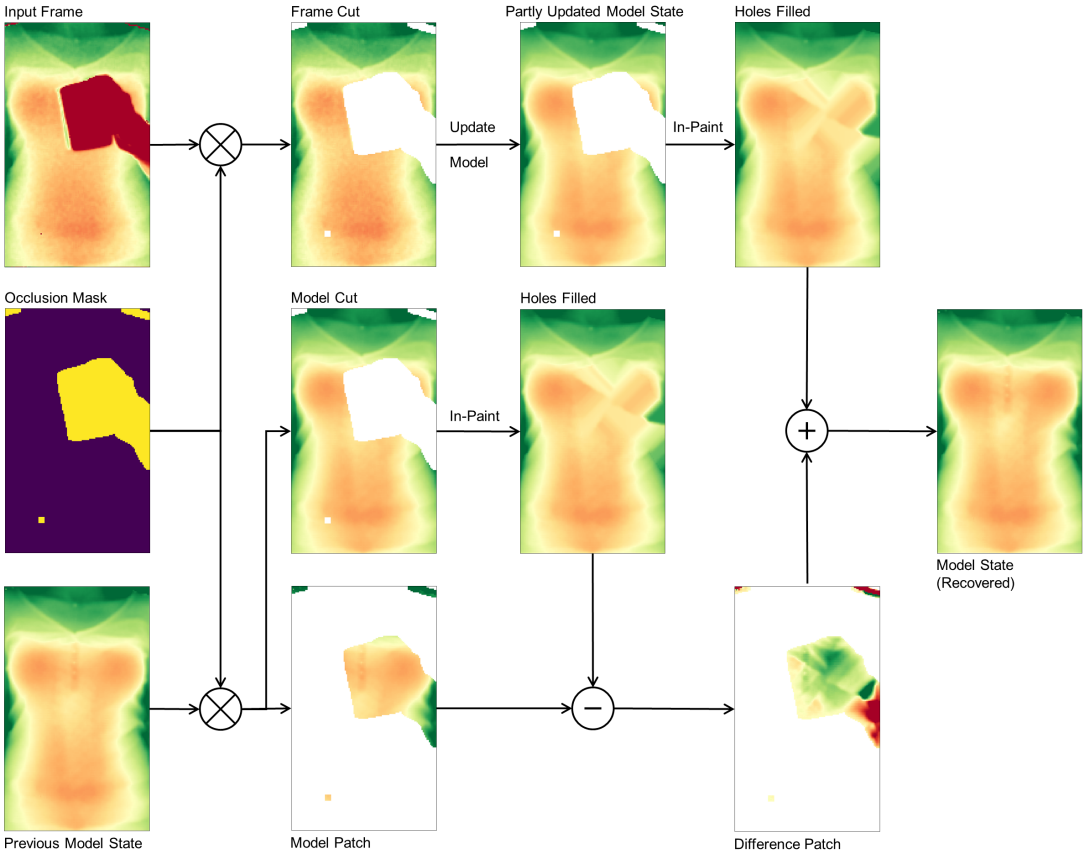

Figure 5.4: Detailed example of the occlusion recovery process: The depth *Input Frame*, *Occlusion Mask*, and *Previous Model State* are used. The occluded area is masked out from the *Input Frame* and the *Previous Model State* to yield the *Frame Cut*, the *Model Cut*, and the *Model Patch*. The non-zero pixels from the *Frame Cut* are fed into the model update routine in order to yield a *Partly Updated Model State*. The holes in the *Partly Updated Model State* and the *Model Cut* then are in-painted using the same technique. Finally, the difference of the *Model Patch* and the in-painted parts of the *Model Cut* are added to the in-painted area of the *Partly Updated Model State*, yielding the recovered *Model State*. First and second order derivatives of the model are updated accordingly.

5.2.3 *Occlusion Recovery*

With the occlusion mask in place, the occluded pixels in the input frame can be identified and, after fitting the candidate windows for the torso window, the model supplies the depth information for the occluded area. After removing all occlusion pixels from the input frame, the remaining depth pixels are fed into the model update routine to yield a partly updated model state with a hole at occluded regions. From the input frame alone, it is not clear how to recover the torso surface. So, in a first step, normalized convolution [84] with a Gaussian kernel is used to in-paint the unknown area. The previous model state, again using the occlusion mask, is separated in two cuts: The model cut describing the valid torso regions, and the model patch describing the occluded torso regions. The hole in the model cut gets in-painted with exactly the same method and parameters as the partly updated model state was filled before.

Subtracting the model patch from the in-painted model cut yields a difference patch that describes the torso surface details. This difference patch now in the final step is added to the partly updated model to recover the surface details of the occluded parts. The first and second order time derivatives of the model eventually have to be updated by feeding the state estimate in the corresponding equations in (5.4).

The occlusion recovery plays an important role in the proposed method as it keeps surface details and helps during the occlusion masking of successive frames to not accidentally mask out parts of the torso. The overall process again is depicted in Figure 5.4.

5.2.4 *Adaptive Torso Model*

It is assumed that users are facing the depth camera and thus users' torso regions will be visible to the depth sensor. Under this assumption, the model becomes a fixed size depth image tracking the torso, along with its first and second order time derivatives (Figure 5.5). The model parameters are updated at each time step when a new depth frame arrives. With the help of the model, the next frame at time t can be predicted by applying (5.3) to each depth pixel x_{t-1} at time $t-1$ independently:

$$x_t = x_{t-1} + \dot{x}_{t-1} + 0.5 \cdot \ddot{x}_{t-1} \tag{5.3}$$

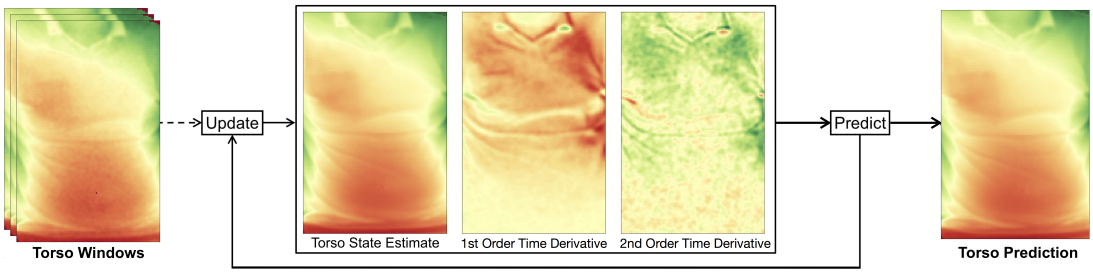

Figure 5.5: This approach builds an adaptive model for the user's torso appearance. From depth input frames, *torso windows* are selected over time that, after occlusion recovery, update the model consisting of a torso state estimate, its first order time derivative, and its second order time derivative. From these, the model can build a torso prediction for the next input frame.

5.2.4.1 *Initialization of the Torso Model*

To initialize the model, a bounding box around the user's torso has to be selected in the input frame. The detected body landmarks and joint positions of the person are used to locate the torso (detailed in Section 5.2.1). The pixels in the model's current state estimate are set to the depth values of the selected region in the very first depth frame. This may contain zero valued depth pixels, so-called holes, which are caused by shadowing, occasional salt and pepper noise due to defect pixels, or reflections from certain materials. For the initialization, these holes are filled with the use of a median filter, whereas afterwards, the model will fill the holes during the occlusion recovery stage as described in Section 5.2.3. The model's first and second order derivatives are initially set to zero and converge after several frames. The model size,

defined by the width and height, is chosen sufficiently large to be able to contain the user's torso. Since multiple rescaling operations need to be performed if the size of the model is allowed to change, this is simplified in this approach by assuming that a person's movement perpendicular to the camera is relatively small.

5.2.4.2 *Model Update with Time Domain Filtering*

Each depth pixel comprises normally distributed noise that increases with distance. Since observed objects or subjects can arbitrarily move in any direction at any random velocity, the noise can not simply be averaged out without obtaining traces from the averaging process on moving objects. For this reason, and since only the depth measurements of the user's torso are of interest, a bounding box is tracked over the X and Y positions of a user's torso, as stated in Section 5.2.1. Within this bounding box, the torso's relative pixel positions along the X- and Y-axis can be assumed to be constant. Consequently, only the motion along the remaining depth axis (or Z-axis) remains and, under the assumption that the user does not move towards the camera, noise can effectively be reduced by a pixel-wise low-pass filter.

In order to achieve a real-time performance, a recursive implementation with a small overhead is used for this filter. Furthermore, the filter needs to follow the signal closely, i.e., with a small delay with little overshoot or damping, as the resulting filtered model is used to detect potential occlusions. For this reason, the double exponential filter is adapted to react faster to an input signal while having less overshoot. Its mathematical definition is stated in Equation (5.4). This behavior is achieved by incorporating the first order time derivative $\dot{x}_{t-1}$ as well as the second order time derivative $\ddot{x}_{t-1}$ in the prediction of the state update x_t from the measurement $x_{t,meas}$ at frame t. The computation of the second order time derivative $\ddot{x}_t$ incorporates the difference of the previous and current velocity approximation $\Delta\dot{x}_t$ as well as a damping factor d to reduce the typical overshoot of the double exponential filter. The filtering equations (5.4) are applied to each pixel separately to yield a smooth state estimate over time. The model then is recursively computed as follows:

$$\begin{aligned} x_t &= \alpha \cdot (x_{t-1} + \dot{x}_{t-1} + 0.5 \cdot \ddot{x}_{t-1}) + (1-\alpha) \cdot (x_{i,meas}) \\ \dot{x}_t &= \beta \cdot (\dot{x}_{t-1}) + (1-\beta) \cdot (x_t - x_{t-1}) \\ \ddot{x}_t &= \gamma \cdot (\ddot{x}_{t-1}) + (1-\gamma) \cdot (\underbrace{(x_t - x_{t-1}) - \dot{x}_{t-1}}_{\Delta\dot{x}_t} - \underbrace{d \cdot \dot{x}_{t-1}}_{damping}) \end{aligned} \tag{5.4}$$

5.2.5 *Extraction of Respiration Signal*

Once the torso window is defined, several approaches to extract a respiration signal from the resulting depth data have been suggested. Simple, yet very effective approaches take for each depth frame the mean of a predefined torso region (for example a window of the chest or a part thereof), and thus track the movement of the chest plane towards and away from the depth camera. This tends to work well for all torso regions that show sufficient respiration motions. It does require the user to keep still, however, as small movements in the range of a few centimeters or even below will affect the measurement. While lying down or sitting on a chair

with sufficient back support, this has proven a straightforward easy task. If the user is standing, though, the torso tends to be subject to a much larger amount of movement. This section proposes a new method that relies on observing both (1) torso regions within the torso window that correlate with breathing motion and (2) torso regions that are barely affected by breathing motion but still are affected by similar torso movements.

Determining the body movement that is not interfered by the breathing motion is not an easy task. The limbs, although not being affected heavily by respiration, do not represent the body movement due to their capability of independent movements. The torso, on the other hand, is heavily affected by respiration. This ranges from the up and down movement of the shoulders, the expansion and contraction of chest and abdomen, down to motion at the hip e.g. caused by a belt that moves during abdominal breathing. Furthermore, clothing plays an important role. A loose dress, for instance, could be hanging from the chest covering the abdomen, leading to breathing movement across the entire torso. Clothing in general leads to varying surface deformations during breathing or movement and it is hard to predict the exact source of the deformation of a garment. These findings render almost all torso regions unsuitable for detecting the body movement as they either are affected by respiration or comprise unpredictable or unrelated motion due to surface deformation.

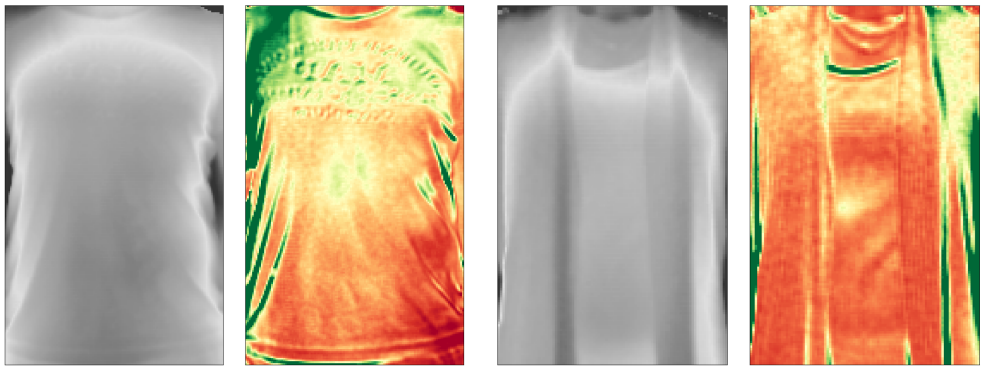

Figure 5.6: Torso surface (left images) and variance of a 12 s time window (right images) of two persons. Red: Low variance; Green: High variance. The throat area shows low variance while the remaining area is highly influenced by clothing in an unpredictable way.

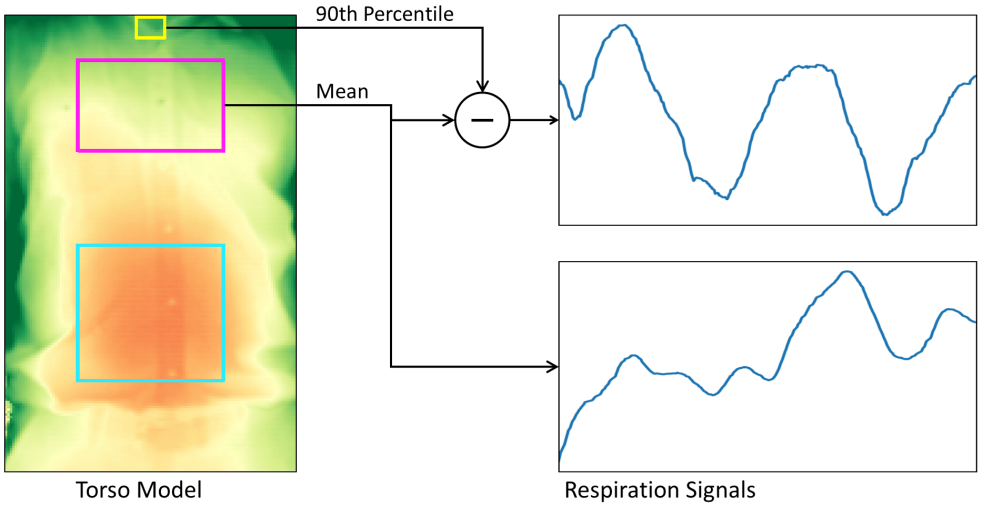

Figure 5.7: The estimation of the respiration signal uses defined areas in the torso window. For the chest region, the result (top right) is a clean respiration signal when subtracting the 90th percentile of the depth values for a small region around the throat and the mean depth measurement over the chest region. In contrast, the mean depth measurements over the chest alone are highly influenced by motion artifacts (bottom right).

Figure 5.6 visualizes the variance of the torso from two persons over the course of 12 s with negligible body movements. Low variances are depicted in red and high variances in green. In preliminary experiments, the throat area was found to be a relatively stable region that barely shows movement caused by chest expansion during breathing. In indoor environments, it furthermore tends to be left uncovered from a scarf or a tight jacket. The throat, however, may partly be occluded by a collar, which causes, due to breathing motion, a significantly higher variance than the bare throat. By only considering the furthest points from the depth camera at a certain region around the throat, the effects of the closer points of a moving collar can be minimized. The 90th percentile of that region has shown good results and was chosen to become the measure of the body movement.

By combining the above observations, the described approach delivers a motion-robust respiration signal by taking the difference of the mean of a highly breathing-affected area, such as the chest, abdomen, or the entire torso, and some sort of a maximum value (the 90th percentile) of a minimally affected area, for which the throat is selected as a good candidate. The affected area is defined with a margin of 20% the window size to the left and right (see [79] for the benefits of a slightly smaller window) and the according vertical position and extent as given by the joints as shown in Figure 5.7. The respiration signal is extracted from the torso model, including regions that currently are subject to occlusion recovery. Simply leaving these regions out would lead to significant signal distortions due to the torso's uneven surface structure: When an occluding entity moves through the frame, step by step, different torso regions become visible that all show a different elevation. Furthermore, in case of full occlusion of the window, no signal could be obtained at all.

5.3 STUDY DESIGN

To validate the proposed method under realistic circumstances and to evaluate the influence of various parameters on depth-based respiration estimation in general, two different studies are performed. The first study, namely the *validation study*, is designed to assess how well the proposed method compares to the actual respiration signal from a chest-worn respiration sensor and the second study, namely the *systematic parameter evaluation*, tests the most common state-of-the-art methods and the proposed method against a range of different input parameters, for instance the observed torso region or the distance to the camera. Both studies are performed with a common setup and in the same environment, which is described in Section 5.3.1.1. Their specific details then are described in Sections 5.3.2 and 5.3.3, respectively.

5.3.1 *Dataset*

5.3.1.1 *Setup and Environment*

The sensor setup consists of a Kinect v2 RGB-D camera and an auxiliary display to show a breathing visualization or a video to the study participants. The depth sensor is fixed to the height of 1.40 m for all recordings and recording was done in a well-lit indoors environment where two adjacent walls with large windows along the entire length of the walls cause challenging lighting conditions. The orientation of the camera was fixed at an angle of 25° towards the floor for sessions where participants are sitting, and at an angle of 0° when participants are standing. This ensures that the participants' entire torso is visible in all depth frames, especially at small distances. Different distances in steps of 1 m are marked on the floor and range from 1 m to 4 m. For the recordings, each participant is asked to position him- or herself comfortably in front of the sensor setup at a specified distance as marked on the floor and such that he or she faces both, the depth sensor and the display. A capturing tool records the raw depth frames and the respective body joint estimates as given from the Kinect SDK and stores the data of each session in a separate file. Figure 5.8 shows some examples of the depth data from a distance of 2 meters for all participants while sitting, while standing upright, and while holding a cup in front of the torso and performing drinking gestures. Overall, more than 11 hours of respiration data were recorded, with about 1.5 hours for the validation study and over 9.5 hours for the systematic parameter evaluation.

5.3.1.2 *Study Participants*

For the experiments and validation of the proposed method, 24 participants were recruited, 17 of them male and 7 of them female, and aged between 22 and 57 years old. Participants were recruited locally and were not diagnosed with respiratory illnesses. Each participant was beforehand shown the depth imaging equipment and was briefed on the study goals and the research questions. All participants were instructed to wear their regular indoors clothing during recordings, ranging from tops, T-shirts, sleeved shirts, collared shirts, sweatshirts to woollen pullovers and hoodies.

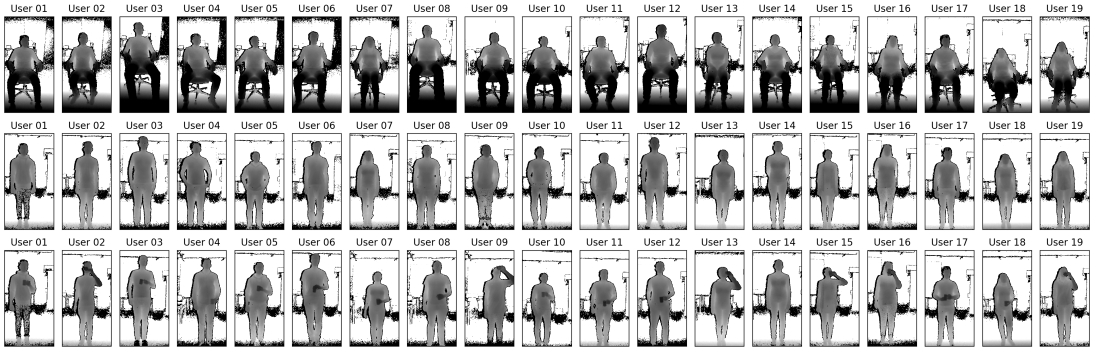

Figure 5.8: Exemplary depth data from the subset of 19 users that participated in the systematic parameter study, each taken at a distance of 2 meters, while sitting (top row), standing (mid row), and while holding a cup and performing drinking gestures (bottom row). During the systematic parameter study participants were asked to sit or stand in front of a depth sensor and follow a breathing visualization while being recorded for 20 sessions, each at 4 different distances from the depth camera, two different paced breathing rates, and for the three conditions sitting, standing, and standing with occlusion, leading to regular occlusions of the torso. In total, 24 study participants were recorded for both studies, with 17 of them male and 7 of them female, and all participants aged between 22 and 57 years old. Overall, 422 unique recordings with a length of over 11 hours were recorded.

Out of the 24 participants, 14 (10 male, 4 female) volunteered for the validation study, and 19 (12 male, 7 female) took part in the systematic parameter evaluation.

5.3.2 *Study Protocol for the Validation Study*

To validate the proposed method, it is compared to the Vernier GDX-RB respiration belt. The respiration belt contains a force sensor tied to a strap, which is to be worn around the person's chest. The measured force is a proportional measure for the chest expansion during breathing and thus is well suited for the verification of the proposed method. It samples at a rate of 20 Hz and transmits the data via Bluetooth Low Energy (BLE) to the same PC workstation that also receives the depth frames.

The validation study is conducted with a set of parameters that are particularly challenging for the extraction of the respiratory rate. Recording hereby lasted at least 120 s, with one at a paced breathing rate, and the two others at natural breathing rates. The 14 study participants were instructed to stand in front of the sensor setup in an upright position at a distance of 3 m while wearing the respiration belt. This setting is challenging to the proposed method due to the distance and due to the standing position that introduces motion artifacts as described in Sections 5.5.4 and 5.5.5, where the influence of the user's posture and distance to the camera are studied. The belt was worn directly under the armpits, following the instructions in the user manual. Furthermore, the belt's sensing device due to its form factor is aligned to the side of the body below the left arm and is worn underneath the clothing to not interfere with the measurements of the depth camera. In total, three different recordings are captured, each comprising a different activity and breathing pattern:

- A paced-breathing *Meditating* recording, where participants were shown a paced breathing display of a growing and shrinking circle with instructions to breathe in and out at a frequency of 0.25 Hz or 15 breaths per minute (bpm). This is a common target breathing rate for meditation, whereas a higher-paced breathing experiment would come with its own risks for the study participants.
- A *Relaxing* recording, while the display is showing a video of landscapes with relaxing background music, to entice a person-dependent slow breathing rate during a relaxing activity.
- A *Post Exercise Recovering* recording, where participants were monitored after strenuous exercise comprising running down and up 12 flights of stairs, and the display showing the aforementioned relaxation videos. This exercise was chosen to heighten the respiratory rate of the study participants. This recording comprises a high variance of the respiratory rate and many random movements from the user breathing heavily, especially in the beginning. Thus, it is considered even more challenging to the proposed method.

Furthermore, participants were asked to refrain from moving their arms a lot, as the tightening of the breast muscles is known to introduce motion artifacts in the respiration belt's sensor data. The impact of self-occlusions are assessed in the systematic parameter evaluation. In total, the dataset of the validation study comprises 42 unique recordings with about 1.5 hours of such respiration data.

5.3.3 *Study Protocol for the Systematic Parameter Evaluation*

The performance of state-of-the-art methods, and how they compare to each other under a series of variable settings, including different distances to the depth camera, different breathing rates, different user postures, and using a variety of study participants, to date, remains unknown. The goal of this study therefore is to evaluate all methods on a common dataset and with expressive performance measures. This section presents the details and recording parameters of the dataset that will be evaluated in Section 5.5.

The 19 participants were told to sit through 20 recording sessions, each lasting for about 5 minutes and interspersed with short 5 minute breaks, with each recording comprising at least 90 seconds of respiration data. The recordings are split into three different conditions:

- In a *sitting* condition, participants were asked to sit in an adjustable office chair in front of the depth sensor. The height of the chair was fixed to 0.5 m, but its back support could be reclined and did not need to be used (i.e., participants could lean back or not, as they preferred). To fix the distances between chair and depth camera, markers were taped to the floor to define the exact positions where the chair had to be placed. Participant were asked to face the depth camera and to keep the arms away from the chest area (e.g., on the chair's armrests) such that the participant's upper body was fully visible to the depth sensor.

- In a *standing* condition, the participants were instructed to stand in an upright position following the same rule as in the first session, i.e. to keep their arms away from the torso region. The goal of this session is to observe the torso's motion while the observed person is standing relatively still, but does not have the support of a chair's seating and back surfaces. Having to stand upright for several minutes tends to introduce a range of motions that are unrelated to the breathing movements of the torso region; Some participants did move their arms in different positions during the recordings (for instance, switching between hands on the back and hands in the pockets) or repositioned themselves to a more comfortable posture, making it potentially challenging to extract a respiration signal from these data.
- In an *occlusion* condition, frequent occlusions were introduced by instructing the participants to hold a cup of tea in front of their torso while standing upright. At the start of the session, participants were recorded for 20 seconds while holding their cup away from the torso. For the remainder of the session, participants were instructed to occlude their stomach and chest regions with the cup by performing drinking gestures. Such self-occlusions also occur when gesticulating, but the drinking gestures were found to be particularly challenging due to their relatively slower speeds of execution and the larger, additional occlusion of an in-hand object. Participants were not required to hold the cup in a particular hand and some participants moved the cup with both hands at the same time to the mouth.

For each participant, these conditions were recorded at distances of 1, 2, 3, and 4 meters between participant and depth camera. For the sitting and standing conditions, the recordings were repeated at two respiratory rates of 10 breaths per minute (bpm) (0.17 Hz) and 15 bpm (0.25 Hz), each. These are obtained through paced breathing. The occlusion condition was recorded with a respiratory rate of 10 bpm (0.17 Hz). For the paced breathing, participants were asked to adhere to a paced breathing visualization shown on the display. The intention is to guide participants' respiration at a stable rate to make the recordings independent of user specific breathing behaviours, such that it does not interfere with the influence of the different other parameters and is more comparable to these parameters. As a reference, the normal respiratory rate of an adult lies between 12 bpm and 20 bpm [29]. Participants did not wear any sensors to exclude effects on the breathing behaviour, for instance due to distraction. Ground truth is obtained from the respective settings in the paced breathing tool. The recording was started after about two minutes, to give the respective participant a chance to adapt his or her respiratory rate to the given target frequency. Overall, the dataset comprises 380 unique recordings with over 9.5 hours of such respiration data.

5.3.4 *Performance Measures*

Overall, four different performance measures are used: The accuracy, the error, the correlation to the ground truth, and the Signal-to-Noise Ratio (SNR). The accuracy describes how accurate the respiratory rate can be computed from the breathing

signal as compared to the ground truth, the error describes, how far the respiratory rate is off from the ground truth, the correlation describes how similar the signal is to the ground truth, and the SNR describes the quality of the signal, i.e. how well the breathing signal stands out of the noise and thus how well it can correctly be extracted.

5.3.4.1 *Accuracy*

For the computation of the accuracy, the breathing signal is shifted to frequency domain with the Fast Fourier Transform (FFT) by using a moving window approach. The moving window has the length l and is moved over the signal with the step size s, splitting the signal up into different equally sized segments. These segments will have a certain overlap that can be defined by both windowing parameters l and s. If the dominant frequency within the range of 0.1 Hz (6 breaths/minute) to 1.5 Hz (90 breaths/minute) of such a segment is equal to the ground truth frequency, this segment is considered a correct estimate. The number of correct estimates divided by the overall number of segments of a single session's respiration signal for a given algorithm is the average accuracy for this session (user, distance, etc.) and algorithm. Its computation formally is stated in equations (5.5) and (5.6).

$$acc(x, \omega_{ref}) = \begin{cases} 1 & \text{if } \underset{0.1<\frac{\omega}{2\pi}<1.5}{\arg\max} \left(\mathcal{F}\{x\}(\omega)\right) = \omega_{ref} \\ 0 & else \end{cases} \tag{5.5}$$

$$Accuracy(x) = \frac{1}{N} \sum_{i=0}^{N} acc([x_{i \cdot s}, x_{i \cdot s + l}], \omega_{ref}) \ , \ x = x_0 ... x_n \tag{5.6}$$

Due to the frequency binning of the FFT, the window length is a crucial parameter for the accuracy computation. A narrow window length yields a good time resolution, providing many segments to test the signal against the ground truth, but cannot provide a fine frequency resolution as a broad spectrum of frequencies will fall into the same frequency bin. This effectively lowers the precision of the accuracy measure since this whole spectrum will be considered a correct estimate. As respiration usually comprises a considerably low frequency, a rather large window size is required to resolve these low frequencies. To yield a precision of one breath per minute, a window covering 60 seconds of data is required. A wide window length on the other hand generates fewer signal segments, effectively reducing the resolution of the accuracy measure. Furthermore, due to their length, signal distortions or short periods of frequency deviations may be shadowed or cause the entire segment to fail the test against the ground truth.

For the validation study, a window length of $l = 40$ seconds and a step size of $s = 10$ seconds is chosen. To reduce frequency leakage, furthermore, a Hann-window is applied to those segments. For the systematic parameter evaluation, a window length of $l = 48$ seconds and a step size of $s = 6$ seconds is chosen. Here, no windowing function is applied since the chosen window length fits both fixed respiratory rates and no frequency leakage is expected. The rationale behind the choosing of the parameters are explained in more detail in Sections 5.4.2 and 5.5, respectively.

5.3.4.2 *Error*

To make the accuracy measure more expressive, a measure for the frequency estimation error is introduced. It tells how far the estimated respiratory rate is off from the ground truth frequency. Hereby, the frequency spectrum as obtained during the accuracy measurement is used. In a first step, the frequency resolution locally is increased by interpolating the dominant frequencies using Quinn's second estimator. The difference of the refined, more precise dominant frequency and the ground truth frequency then becomes the estimation error as again formally defined in Equation (5.7). The error first is computed separately on each single window as used in the accuracy computation, and the different windows' errors afterwards are averaged to yield the mean error of the whole sequence of a single recording. The error is defined as:

$$\text{Error}(x, \omega_{ref}) = \frac{1}{N} \sum_{i=0}^{N} | \underset{0.1 < \frac{\omega}{2\pi} < 1.5}{\arg\max} (\text{Quinn}_2(\mathcal{F}\{[x_{i \cdot s}, x_{i \cdot s+1}]\}(\omega))) - \omega_{ref}| \quad (5.7)$$

5.3.4.3 *Pearson Correlation Coefficient*

The similarity of the proposed method's estimated breathing time series to a given ground truth signal is assessed by computing their Pearson Correlation Coefficient (PCC) as given in (5.8):

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (5.8)$$

The respective confidence intervals CI are computed with the Fisher transformation as stated in (5.9). The standard error SE, given the sample size n, is computed with (5.10), and the z-score z for a desired confidence interval of $(1-\alpha)\%$ is obtained from the standard normal cumulative distribution function.

$$CI = \tanh(\text{arctanh}(r_{xy} \pm z_{\alpha/2} \cdot SE)) \quad (5.9)$$

$$SE = \frac{1}{\sqrt{n-3}} \quad (5.10)$$

Since an expressive quality measure of the overall signal is of interest and to ensure a high confidence of the PCC, it is computed on the whole, fixed length of the signal. From Equation (5.9) follows that due to the several thousand samples of any of the signals, even the 99% confidence intervals are narrow and only a small fraction apart from the computed PCC. Consequently, all confidence intervals given during the evaluation refer to the 99% confidence intervals. The ground truth signal either is obtained from the body-worn respiration belt or it is given as a sine signal obtained from the frequency settings of the paced respiration setup.

5.3.4.4 *Signal-to-Noise Ratio*

The signal quality is measured in terms of the Signal-to-Noise Ratio (SNR) as defined in Equation (5.11). A higher SNR value means that the respiratory signal more significantly stands out from the noise and therefore is easier to extract from the data. The SNR also is computed on the signal as a whole and is given as:

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \cdot \log_{10} \left(\frac{P_{Signal}}{P_{Noise}} \right) \tag{5.11}$$

5.4 VALIDATION STUDY OF THE PROPOSED METHOD

The goal of the validation study is to validate the proposed method by comparing its performance to that of a commercial wearable respiration belt. Three different activities were taken as different conditions: Following a paced-breathing meditation, relaxing, and recovering after a sports exercise. Before processing, a Butterworth 5th-order band-pass filter is applied to both signals. The lower and upper cut-off frequencies are set to the same range that is used to find the dominant frequency (see Section 5.3.4), namely to 0.1 Hz (6 breaths/min) and 1.5 Hz (90 breaths/min). The filter is applied both in forward and backward direction to minimize transients at the start and end of the signal. The band-pass filtering allows a better comparison of the proposed method to the respiration belt as it removes constant or non-linear offsets from the data and reduces high frequency noise while preserving the breathing-relevant range of frequencies. Filtering, however, is not required in general as shown in Section 5.5, where good evaluation results are obtained from the proposed method without filtering.

Before stepping into the quantitative evaluation of the proposed method against a respiration belt, first a visual comparison between both is conducted in the next section.

5.4.1 *Visual Inspection*

The visual inspection aims at examining the differences of the signals of the proposed method to the ground truth data of the respiration belt. Two example plots of the band-pass filtered respiration signals are depicted in Figure 5.9. Both plots are taken from the post exercise setting and comprise a high dynamic frequency range with a fast respiratory rate at the beginning of the recording (left part of the plots) that decreases over time to a more nominal breathing rate (right part of the plots). Depicted in Figure 5.9's top plot are the data from a recording where this works well, with the resulting performance for the PCC of 0.95 and an accuracy of 100%. A second, more challenging recording is plotted below in Figure 5.9, with a PCC of 0.26 and an accuracy of 22%, due to the user moving more during the recording.

At first sight, the main differences lie in signal amplitude and quality. While the best case signal comprises a high amplitude and only small deviations from the ground truth, the worst case signal contains much smaller amplitudes and many peaks and deviations over the whole spectrum, but especially at higher frequencies. The varying amplitude is caused by the low-pass behaviour of the model. More

severe, however, are the peaks. These stem from the user bending forward while heavily breathing, thus covering his throat with his head. The proposed method in this case no longer is able to fully reconstruct the relatively small reference region and motion artifacts enter the signal. The model is able to recover after such an occlusion event and the signal follows the ground truth well until the next occlusion occurs. The smaller deviations at lower frequencies (right part of the plot) are caused by strong movements that could not fully be removed from the signal, but, to a certain extent, still contain the respiration signal.

In general, in terms of frequency, the proposed method matches the ground truth well in both cases, but the many occasional peaks in one signal hinder it in estimating the correct respiratory rate.

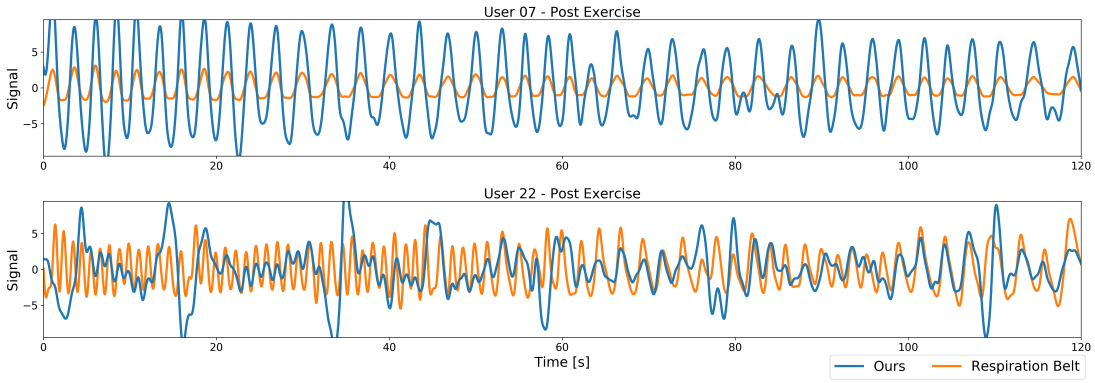

Figure 5.9: Comparison plots between the output of the chest-worn respiration belt (in orange) and the output of the proposed method (in blue), for the post exercise condition (going from a fast respiratory rate in the beginning to a more nominal one over time). The top segment has a PCC of 0.95 and an accuracy of 100% and both signals match well in terms of frequency. The bottom segment has a PCC of 0.26 and an accuracy of 22%, with the larger peaks on the left in the output due to the user occasionally tilting the torso forward and occluding the throat region with the head.

5.4.2 *Quantitative Evaluation*

After the visual inspection, the proposed method's performance is quantitatively investigated for all study participants and respiration patterns. Figure 5.10 depicts the PCC, the accuracy, and the error of all study sessions, for each user separately. For the accuracy and error measurements, a FFT window length of 40 s and a step size of 10 s is used. To reduce frequency leakage, moreover, a Hann-window is applied to those segments.

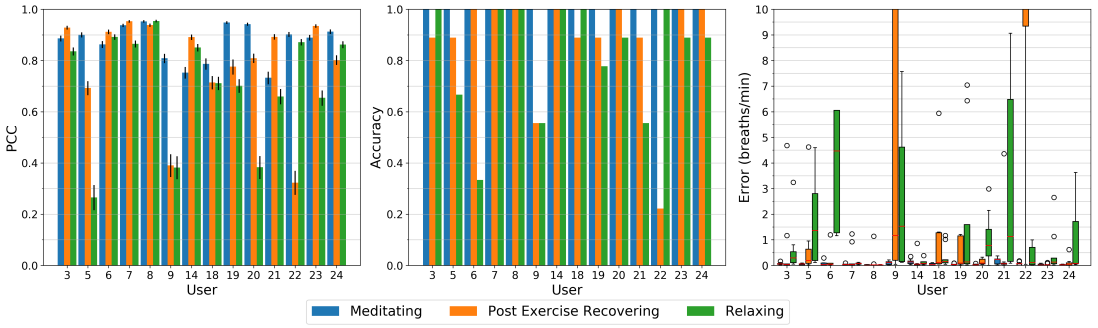

Figure 5.10: **Left**: The Pearson Correlation Coefficient between the data from the Vernier respiration belt and the proposed system, for all users individually. The bars depict the means, while the black bars indicate the 99% confidence intervals. Data from paced breathing (15 breaths per minute) tends to result in high correlation between the proposed method's prediction and the belt's output. Relaxed and post exercise breathing tends to perform slightly worse. **Middle**: The accuracy of the proposed system compared to the respiration belt. **Right**: The error of the estimated respiratory rate compared to the respiratory rate from the respiration belt. Poor performance for users 9 and 22 stem mostly from larger movements during the recording.

Paced. The high-paced breathing meditation has an accuracy of 100% for all participants and close to zero errors. All signals show high correlation to the respiration belt with a minimum PCC of about 0.75 for users 14 and 21, up to a value of 0.9 and above for about half of the users.

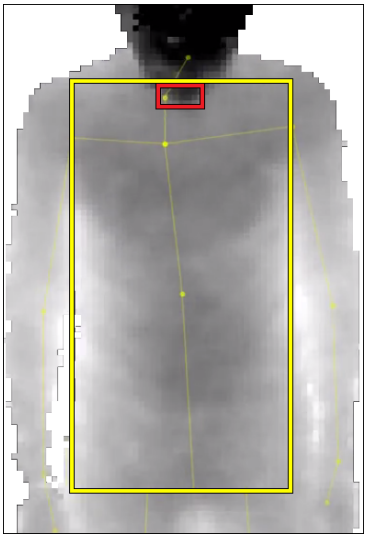

Figure 5.11: User 9 bending its head forwards due to heavy breathing, thereby occluding the reference region at the throat (marked in red).

Post Exercise Recovering. The post exercise respiration shows remarkably high errors for users 9 and 22. All other users either have almost zero errors or errors in the range of about 0.5 to 1.5 breaths per minute as in the case of users 5, 18, and 19. User 9 and 22 also have significantly lower correlation coefficients below 0.4 as compared to other users, mainly caused by occlusion events where both, due to heavy breathing, lower their head and occlude the proposed algorithm's reference region at the throat as shown in Figure 5.11. In the case of user 22, as described above, the occlusion recovery cannot fully restore that region, leading to peaks in the signal as illustrated in Figure 5.9 (bottom). For user 9, this region sometimes can be recovered to a certain extent, but the respiration signal is significantly raised or lowered during that time. These events have similarities in the time domain with a rectangular shape, so harmonics show up in the frequency domain that dominate the respiration frequency.

Relaxed. On relaxed breathing, the proposed method performs worse than on the other respiration patterns. Notably, some users show a low accuracy with values of 65% for user 5, about 55% for users 9 and 21, and below 35% for user 6. These users also show the highest error rates with a median above one breath per minute, especially for user 6 with a median error of 4.5 breaths per minute. Its PCC however is with a value of 0.9 very high which indicates that due to small signal deformations and due to the relatively small window length used for the low breathing rate, other frequency components become more dominant in frequency domain. Users 5, 9, and 20 have a PCC below 0.4 and can be considered to be only weakly correlated to the respiration signal from the chest strap. User 20 on the other hand has a high accuracy of about 90% but still a median error of about 0.7 breaths per minute.

Summary. To conclude, for most users and most regular breathing frequencies, the proposed approach performs close to the commercial body-worn respiration belt and closely matches the respiratory rate. For the changing frequencies in the post exercise sessions and the relaxed sessions, a few users did move a lot more during the experiment, thus causing a significant drop in their recordings' performance. The heaviest impact on the respiration signal was caused by the user lowering its head and occluding the throat region where the reference signal for the body movement is sampled from. In such cases, such as the one illustrated in Figure 5.9 (bottom), multiple breathing cycles are missed and the proposed method under-estimates the breathing rate.

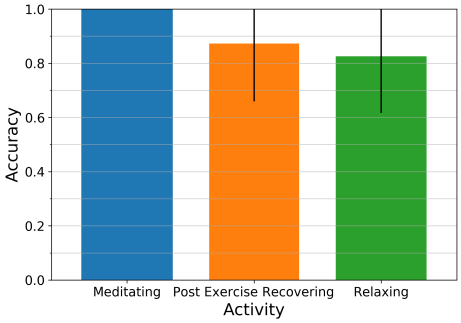

Figure 5.12: Mean and standard deviation of the accuracy as achieved by the proposed method on the activities *Meditating*, *Post Exercise Recovering*, and *Relaxing* across all participants.

Figure 5.12 displays the mean and standard deviation of the accuracy across all users for the proposed method. The proposed method shows for all activities a mean accuracy above 80% and can estimate the respiratory rate of the paced-breathing meditating exercise perfectly with an accuracy of 100%. The two remaining activities, however, show relatively large standard deviations as can be retraced in Figure 5.10 (middle).

As a reference, according to [98], the variability between two successive respiration measurements performed by two different persons in a clinical context (i.e. from doctors or nurses) may account for a difference of up to 6 bpm.

5.5 SYSTEMATIC PARAMETER EVALUATION

In this section, the influence of various parameters on a selection of six different methods, including the proposed approach, are investigated. The methods to be discussed are summarized in Section 5.5.1. The parameters under consideration are the region of interest (chest, abdomen, or the entire torso), the condition (sitting, standing, or standing with occlusions), the distance of the participant to the depth camera

(1 m to 4 m), the user's breathing rate (10 or 15 bpm), and the gender. Additionally, some user dependent observations that were made during the evaluation are summarized at the end of this section. For all accuracy and error evaluations, a fixed FFT window length of 48 seconds is used. It has the advantage that both the 0.17 Hz and 0.25 Hz frequencies from the paced breathing setup can accurately be resolved by a simple rectangular windowing function without frequency leakage occurring at the target frequencies. The window is moved over the signal with a step size of one breathing cycle, i.e. with 6s at 0.17 Hz (10 bpm) and 4s at 0.25 Hz (15 bpm). Overall, the windowing yields a frequency resolution of about 0.02 Hz or 1.2 bpm and 7 or 10 distinct windows to test for the accuracy. The signals furthermore are evaluated on the raw output of the algorithms, i.e. there is no filtering applied to the signals in the following results. An example of the output signals will be given in Figure 5.14 and is discussed in Section 5.5.2.

The sections following the visual inspection then will each focus on a separate parameter and provide a detailed discussion of its influence on the respiration estimation of the different methods. Hereby, the performance measures as introduced in Section 5.3.4 are used to facilitate an objective comparison of the methods.

5.5.1 *Methods Overview*

From the related work, two approaches for depth-based remote respiration estimation can be found that most commonly are used. So, together with the proposed method, in total three distinct approaches can be compiled. These are based on (1) performing a principal component analysis, (2) computing the mean of a certain area that is affected by breathing, and (3) taking the difference of a barely breathing correlated region from the mean of a highly affected region using a torso surface model (the proposed method). From these three distinct approaches, overall six variants are derived for the systematic parameter evaluation. Namely, these methods are the *PCA*, *Mean Raw*, *Median Raw*, *Diff Mean*, *Diff Median*, and *Diff Model*. The named methods and their particular details are described in the following, but first a short introduction is given on how the region is selected that will be used to extract the breathing signal from and that will be common for all methods. The focus hereby lies on an indoors setting where a user is facing a depth camera that also tracks the user's body joints. These body joints, namely the neck, the hip, and both left and right shoulder joint positions as estimated by the Kinect v2 framework are used to define the breathing relevant regions of interest. The hip and shoulder joint positions hereby define the anchor points of the torso window. The torso area itself has a margin of 20% to the left and right side of the window spanned up by the joints and subsequently is subdivided into the chest and abdomen regions. All three regions, the entire torso, the chest, and the abdomen will be examined for their suitability of extracting a respiration signal and thus are sampled from by all methods independently. The neck joint on the other hand only serves as the anchor point for a barely respiration affected reference area at the throat and thus will only be used by difference-based methods that use this region to subtract the motion component from the breathing signal. The six different methods are detailed in the following and a visual overview of all methods as well as the different body regions can be found in Figure 5.13.

PCA. Methods based on performing a Principal Component Analysis (PCA) are a common approach to compute the respiration signal from depth images. As mentioned above, the hip and shoulder joint position estimates are used to find the respective region of interest. Due to the PCA computation requiring a predefined number of pixels, the window's size needs to be fixed to certain extents. The size is given from the shoulder and hip joint positions of the very first frame. The fixed window, however, is free to move and will be anchored on the left shoulder joint position from the respective frame. Fitting the PCA model is done with the first 180 frames, or the first 6 seconds of the capture sequence. The respiration signal afterwards is given from the first component of the PCA model and, for the evaluation, will be computed from all frames, including the first 180 frames. In the following, *PCA* will refer to the method that uses a Principal Component Analysis to extract the respiration signal, not to the Principal Component Analysis itself.

Mean Raw and **Median Raw**. Mean based methods form the majority of the current state of the art. The respiration signal is extracted by, for each frame, computing the mean of all depth values within a given region of interest as defined by the hip and shoulder joints. This region, for each frame, is free to change in size and position, which, in addition to the simple computation of the mean, is a big advantage of this method. The *Median Raw* method basically is the same as the *Mean Raw*, but instead of the mean, computes the median of the given region of interest. The median likely will be more robust than the mean, especially in the case of surface deformation or occlusion. All three methods described so far are likely to be sensitive to motion, occlusion, and window misalignment.

Diff Mean and **Diff Median**. Motion artifacts caused by even small whole body movements, like swaying while standing, can have a significant impact on the respiration signal quality. To overcome such motion artifacts, the proposed difference-based methods try to subtract the motion from the actual respiration movements of the body. They rely on subtracting the signal of a reference area that barely is affected by breathing from a signal of one of the highly breathing correlated regions at the chest, abdomen, or the entire torso. The region around the throat was found to be minimally affected by breathing while serving as a good reference for motion artifacts of the upper body. Both, the *Diff Mean* and the *Diff Median* first compute the mean or the median of a torso region that is highly affected by breathing and this far are identical to the *Mean Raw* or *Median Raw*, respectively. In a second step, the motion reference signal as given by the 90th percentile of the region around the throat is subtracted from the previously computed respiration signal. The region at the throat is determined with the help of the neck and shoulder joints. Both methods are derived from the proposed, model-based method as described in Section 5.2. Their advantage is that they do not need a model, are easy and fast to compute, comprise a mechanism to counteract motion artifacts, and that the window comprising the observed torso region is free to change in size and position from frame to frame.

Diff Model. To compensate for noise, window misalignment, and especially occlusion, the aforementioned method, as proposed in Section 5.2, was introduced. It is able to counteract these issues by low-pass filtering the data, fitting the window to the most reasonable body area, and by detecting and recovering occluded regions with an image in-painting technique. This method computes an internal model of the torso surface area spanning from the throat to the hip and that is based on the currently and previously captured depth images and body joint positions. The model outputs an aligned, occlusion recovered, and noise reduced depth image of the torso that can be used for extracting the respiration signal. To compute the respiration signal, the difference-based approach is used, which subtracts the 90th percentile of the throat region from the mean of the respective torso region. The regions hereby again are defined by the joint positions. This method is more stable against noise, window misalignment, motion artifacts, and occlusion, but also has higher computational complexity and, in the current form, requires a fixed window size for the torso region that needs to be initialized in the beginning.

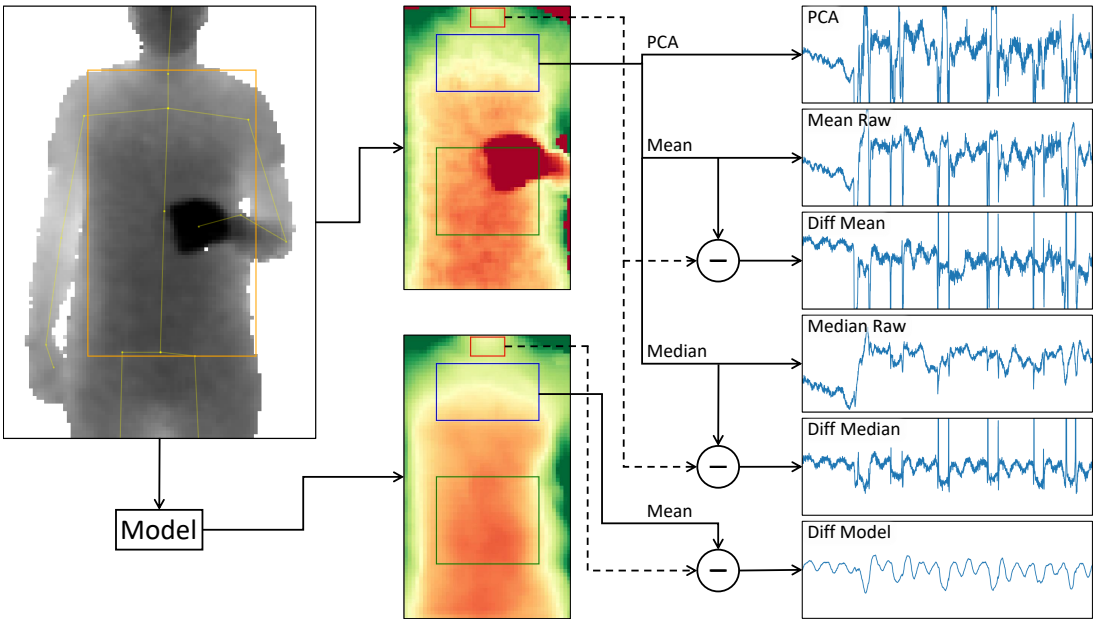

Figure 5.13: Overview of the methods used for the systematic parameter evaluation. The example is taken from a distance of 3 meters, with the user standing upright and performing regular self occlusions with a cup in his hands. The process starts with the camera's depth *Input Frame* and the estimated joint positions of the user (top left). Both are either forwarded to a *Model* as proposed in Section 5.2 (bottom left) to reconstruct the torso surface and find the regions of interest (bottom mid) or the joint positions are used directly to find the regions of interest (top mid). In the latter case, the torso surface is redrawn for comparison purposes to the model output (middle images). The model is able to filter out most of the noise and to recover occluded torso regions. The regions of interest are the throat (red), the chest (blue), the abdomen (green), and the torso (chest and abdomen windows combined, including the region in between both, not drawn in the images). The depth pixel values within the different regions, in this example the values of the chest and, depending on the method, the throat region, are used to compute a single respiration state value. The respiration signal then is given by the history of these values. On the right are the plots of the resulting breathing signals of the different methods. From top to bottom: The signal of the *PCA*, *Mean Raw*, *Diff Mean*, *Median Raw*, *Diff Median*, and the *Diff Model*. The *PCA* uses the first 180 input frames (6 seconds) to compute the principal components, the respiration signal then is computed from the first component of the *PCA* model. The *Mean Raw* and *Median Raw* methods compute the mean or the median of the depth values within the given torso region, for instance the chest as shown here. The *Diff Mean* and *Diff Median* methods on the other hand use the 90th percentile of the throat region depth values as reference for the user movement and subtract it from the respective values obtained by their *Mean Raw* or *Median Raw* counterparts. Their signals contain less distortions stemming from body movements, like swaying. The *Diff Model* method does the same, but computes the difference from the mean of the selected region of the model output. Its breathing signal is much smoother and barely contains distortions or spikes stemming from motion artifacts and occlusion events.

5.5.2 *Visual Inspection*

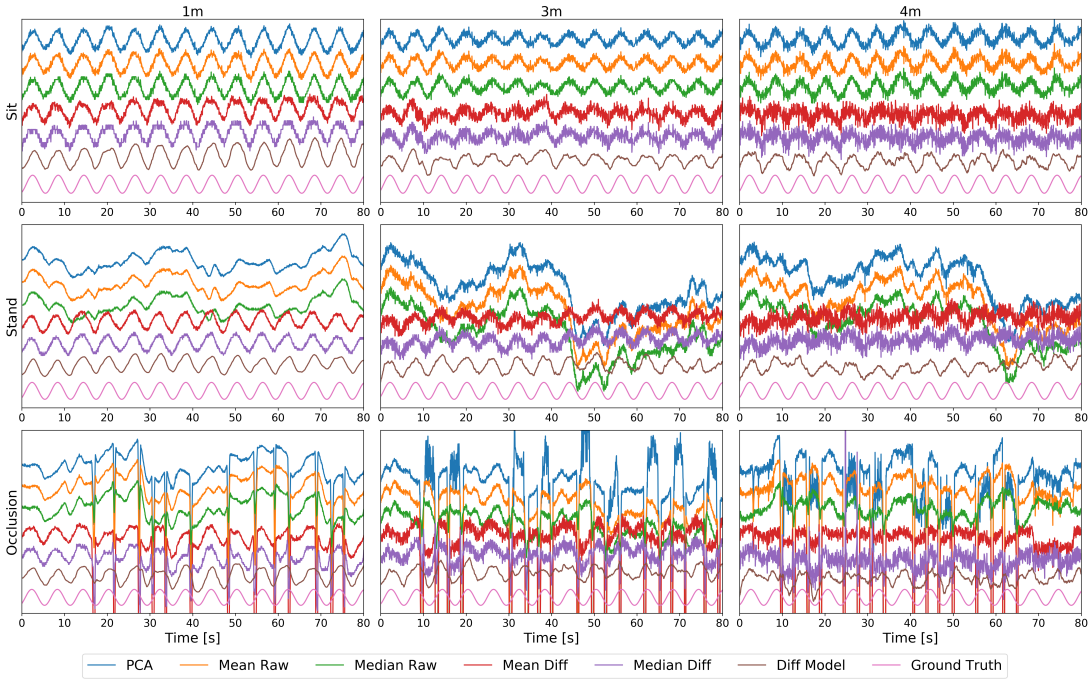

Figure 5.14: The respiration signals from the methods *PCA*, *Mean Raw*, *Median Raw*, *Mean Diff*, *Median Diff*, and *Diff Model*, as well as the *Ground Truth*, obtained from the chest at a distance of 1 m (left), 3 m (middle), and 4 m (right) for the conditions sitting (top), standing (middle), and standing with occlusion (bottom). All signals are zero centered and normalized with respect to the mean and standard deviation of their first 240 frames (8 seconds) and, for better visibility, are stacked vertically in the order as mentioned above. With increasing distance, the noise level of all methods increases. The *Diff Model* hereby is the least affected method and can suppress the noise even at high distances, while the other difference-based methods are affected the most. Standing introduces significant signal distortions that can be seen on the output of non-difference-based methods. Occlusion events, visible as large spikes in the bottom plots, can only be handled by the *Diff Model* method.

Figure 5.14 depicts an example of the signals obtained from the various methods, with the different distances at 1 m, 3 m, and 4 m in the columns, and with the conditions sitting, standing, and standing with occlusion in the rows. For all conditions and methods, with increasing distance an increase of the overall noise level can be observed. Especially the difference-based mean and median methods are strongly affected by noise, since both methods rely on subtracting two noisy signals, which increases their overall noise level. The *Diff Model* method has a built-in low-pass filter to prevent this effect from happening. For this reason, it has the cleanest output signal among all methods, but still shows some smaller distortions at higher distances. The *PCA* method, although only using the strongest component, was not able to separate out the noise from the signal.

While the sitting condition can be managed by all methods, standing introduces small swaying movements, typically in the range of a few centimeters or less, which may introduce severe signal distortions for all non-difference-based methods, as shown in this example. The breathing cycles, to some extent, are still visible in the distorted signals, but other frequency components clearly dominate. The difference-based methods are able to reduce the motion components and are barely affected by them.

The large spikes caused by occlusion events, as seen in the bottom plots, cause even more severe signal distortions and make it difficult to obtain a good signal. The *Diff Model* can internally detect and recover occluded body parts and is the only method that is not or barely affected by occlusion. The median based methods can partially deal with occlusion or at least can limit the spikes to a certain extent.

5.5.3 *The Influence of the Torso Region*

The Influence of the choice of the observed torso region on the detection of the respiration signal yet is unknown for depth-based respiration estimation in general. In the following, thus, the role of the torso region is investigated for all methods as mentioned above, including a comparison of their performance to each other. Figure 5.15 depicts the accuracies, errors, Pearson Correlation Coefficients, and Signal-to-Noise Ratios of the different methods when applied to the chest, abdomen, or the entire torso. All 380 recordings, comprising different users, distances, respiratory rates, and conditions (sitting, standing, and standing with occlusion), are combined in these plots. The observations made here thus show each method's overall performance on the respective region. Furthermore, not a single parameter combination was found that is more beneficial on a different body region other than suggested by these plots. The choice of the window position affects the performance of all other parameter settings in the same or in a similar way. The condition, to some extent, has an influence on the choice of the window position as for instance in the occlusion scenario the abdomen was occluded for longer time periods and more often than the chest. This, however, does not change the observed trend. As a reference, Figure 5.16 in Section. 5.5.4 depicts the influence of all three conditions at the different specific torso regions and thus the influence of a torso region during one of the different conditions can easily be derived as well. Section. 5.5.4, however, provides more details on the influence of the condition, while the current section focuses on the overall performance of the different methods at the different body regions.

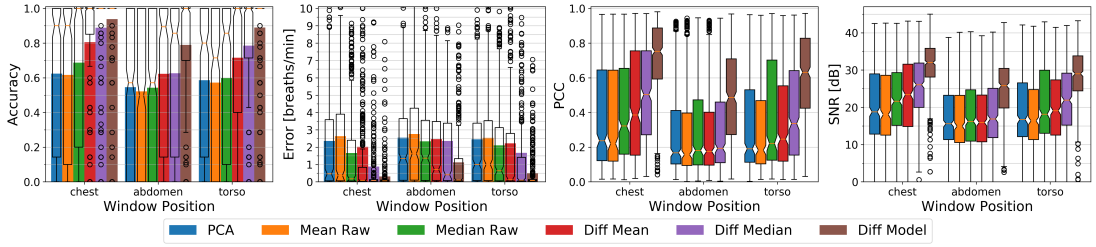

Figure 5.15: From left to right: The respiratory rate accuracy, errors in breath per minute, Pearson Correlation Coefficient w.r.t. the ground truth signal, and Signal-to-Noise Ratio of the different methods. Plots are separated by the chest, abdomen, and torso region with all conditions (sitting, standing upright, and occlusion), distances (1 to 4 meters), and respiratory rates (10 bpm and 15 bpm) combined. Accuracy and error metrics use a FFT window with a length of 48 seconds. The colored bars show the averages, while overlay box plots show median (middle parts) and whiskers marking data within 1.5 Interquartile Range (IQR). All algorithms perform best on the chest region, while the abdomen especially for the difference-based methods causes high performance drops.

Chest. At the chest, the *PCA* and *Mean Raw* methods show with a mean accuracy of about 62% (median 90%) and a mean error of above 2.3 bpm (median about 0.4 bpm) the lowest performance values. The *PCA* hereby performs a little bit better than the *Mean Raw*. The *Median Raw* achieves with a mean accuracy of 69% (median 100%) and a mean error of 1.66 bpm (median 0.17 bpm) slightly better performance values and seems to be more robust than the *PCA* and *Mean Raw* methods (also see Section 5.5.4). The *Diff Mean* likewise is with an average accuracy of 80% (median 100%) and a mean error of 2.0 bpm (median 0.08 bpm) outperformed by its *Diff Median* counterpart, which has an average accuracy of 89% (median 100%) and a mean error of 0.87 bpm (median 0.06 bpm). Both methods clearly benefit from subtracting the motion component obtained from the throat, since without the subtraction, both are identical to the *Mean Raw* or *Median Raw* respectively. The highest performance is achieved by the *Diff Model* method. At the chest, it has a mean accuracy of 94% (median 100%) and a mean error of 0.3 bpm (median 0.06 bpm). The box-plot overlays of the accuracy plots furthermore reveal that the *Diff Median* and the *Diff Model* are able to correctly estimate the respiratory rate of the majority of the 380 samples, except for the outliers marked as circles. There are some differences, however. While the *Diff Model*'s accuracy is only about 5% above that of the *Diff Median*, its mean error is almost three times lower.

In terms of signal quality, the *PCA* and the *Mean Raw* show a median PCC of about 0.22, and a median SNR of about 18 dB. The *Median Raw* achieves with a median PCC of 0.32 and a median SNR of 21.5 dB slightly better values. The *Diff Mean* likewise is with a median PCC of 0.39 and a median SNR of 23 dB outperformed by the *Diff Median* with its median PCC of 0.5 and median SNR of 26 dB. The *Diff Model* achieves with a median PCC of 0.75 and a median SNR of 32 dB a notably higher PCC and SNR than all other methods.

Abdomen. At the abdomen, the *PCA*, *Mean Raw*, and *Median Raw* show with a mean and median accuracy of about 53% and a mean error of about 2.3 bpm to 2.8 bpm (median 1.4 bpm to 1.6 bpm) a similar performance. Compared to the chest, the

Median Raw thus has a higher performance loss than the other two methods. The *Diff Mean* and the *Diff Median* also show a similar performance. Their mean accuracy lies at about 62% (median 85% to 89%) and their mean error at about 2.4 bpm (median 0.4 bpm to 0.7 bpm). Both methods, but especially the *Diff Median*, show the highest loss in performance as compared to the chest. The *Diff Model*'s performance also significantly lowers at the abdomen, but with a mean accuracy of 79% (median 100%) and a mean error of 1.1 bpm (median 0.1 bpm) it still outperforms all other methods. In terms of signal quality, all methods, except for the *Diff Model*, show a median PCC in the range of 0.16 to 0.19 and a median SNR in the range of 15 dB to 17 dB. The *Diff Model* on the other hand has a median PCC of 0.49, and a median SNR of 26 dB.

Torso. The torso region includes both, the chest and the abdomen, and likewise yields intermediate results between both other regions. The difference-based methods hereby again outperform the other methods and end up more favourably than at the abdomen. The *Diff Model* furthermore with a mean accuracy of about 89% only loses about 5% in accuracy as compared to the chest, while both other difference-based methods lose about 10% in accuracy. Also the *Diff Model*'s mean error of about 0.5 breaths per minute is considerably lower and closer to the error at the chest than that of the other methods.

Summary. Overall, the chest region is the optimal choice. It yields, regardless of the method used, the highest accuracy, lowest errors, highest PCCs, and highest SNRs. The abdomen has shown to be the least suitable region for detecting the respiration signal and, in relation to the other regions, marks the lower bound on all performance metrics.

All methods arguably benefit from a larger signal amplitude that, during breathing, stems from a greater expansion of the chest than of the abdomen. Another aspect that needs to be considered is that during the occlusion condition, the abdomen was the body region that was occluded most of the time which further lowers the detectability of the respiration signal.

Comparing the different methods among themselves shows that the *Diff Model* method, regardless of the observed body region, overall is superior to all other methods, followed by the *Diff Median* and the *Diff Mean*. The accuracy box plots furthermore suggest that, except for some outliers, the *Diff Model* as well as the *Diff Median* methods can at the chest optimally estimate the respiratory rate. The weakest methods are the *PCA* and the *Mean Raw*. The difference-based methods, however, comprise larger performance drops at the abdomen or torso than the other methods, which means they are more susceptible to the choice of the body region. The difference-based methods likely perform comparably worse on the abdomen due to the spatial distance of the abdomen to the throat, where the reference region for subtracting the motion components is located. A swaying motion has a larger amplitude at the throat than on the abdomen and additionally the upper body can, to a certain extent, move independently from the lower body, whereas the chest motion can be assumed to be similar to the throat motion.

On all body regions, the *Diff Model* has a notably better signal quality than all other methods. One reason for the *Diff Model*'s higher PCC and SNR is, as suggested by its accuracy and error values, that the true breathing signal can better be estimated

by this method, but this alone does not explain the relatively big difference to the *Diff Median*. The main reason is that the *Diff Model* method uses a low-pass filtering technique and thus is able to model the torso surface with a significantly reduced noise level. From the improved torso surface reconstruction it then can extract a much cleaner respiration signal, consequently leading to higher PCC and SNR values.

Since the chest has been shown to be the most suited region for extracting the respiration signal, the focus will lie on this region in the following sections. Beginning with the influence of the condition, step by step a deeper insight into the specific influence of each single parameter on the overall performance of the different methods will be provided.

5.5.4 *The Influence of the Condition (Sit, Stand, Occlusion)*

The methods proposed in previous works primarily have been evaluated in scenarios where the study participants were lying down or sitting still. In a more realistic scenario, however, the observed person should also be allowed to stand in front of the camera, possibly performing regular self-occlusion gestures. For this reason, this section assesses the performance of the different methods for the three conditions sitting, standing, and standing with self-occlusions by performing drinking gestures with a cup. Figure 5.16 plots the accuracies, errors, PCC values, and Signal-to-Noise Ratios of the different methods against the three mentioned conditions. As stated above, the primary focus lies on the chest, but for completeness, in Figure 5.16 also the evaluation data of the abdomen and the torso is appended. Here, also the influence of the torso region during the different conditions can be derived. Where appropriate, specific findings are highlighted that are strongly influenced by the condition as well as by the observed torso region and thus could not be considered in full detail in the previous section.

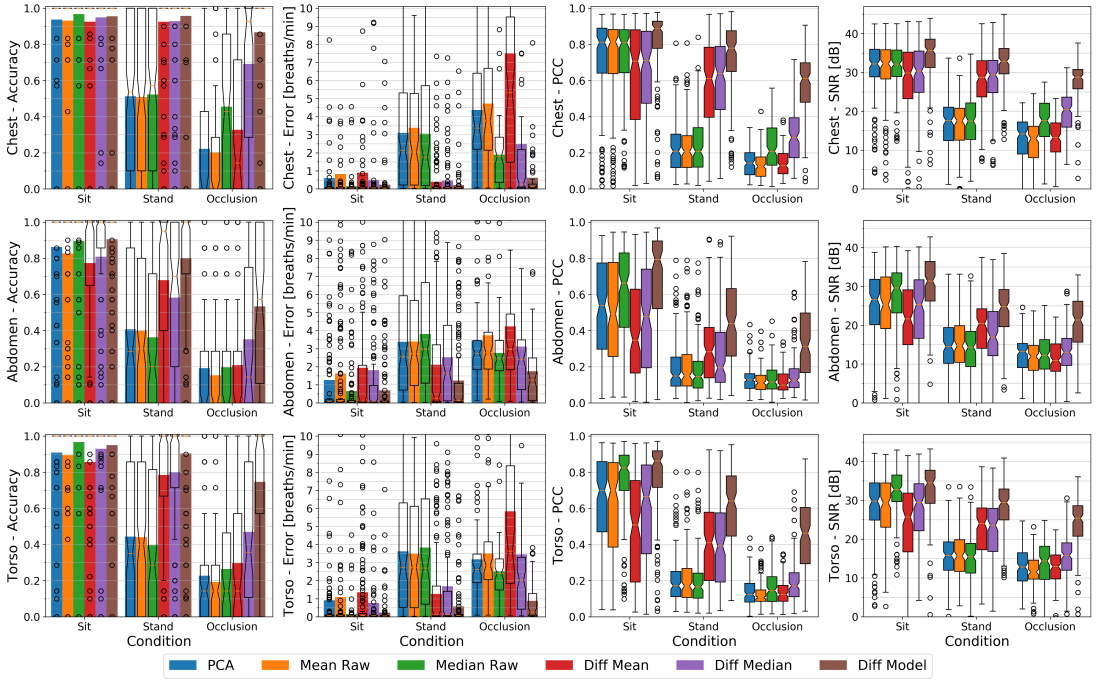

Figure 5.16: From left to right: The respiratory rate accuracy, errors in breath per minute, Pearson Correlation Coefficient w.r.t. the ground truth signal, and Signal-to-Noise Ratio of the different methods. Plots are separated from top to bottom by the chest, abdomen, and torso region and by the sitting, standing, and occlusion condition, with all distances (1 to 4 meters) and respiratory rates (10 bpm and 15 bpm) combined. Accuracy and error metrics use a FFT window with a length of 48 seconds. The colored bars show the averages, while overlay box plots show median (middle parts) and whiskers marking data within 1.5 IQR. Sitting results for all methods in a much clearer signal than standing upright, with standing and occlusion (holding a cup and performing drinking gestures, right measures) performing worse than just standing upright.

Sitting (Chest). Sitting still (or lying down) barely introduces motion artifacts and all methods in previous work have been evaluated for a static scenario like sitting or lying down. So, as expected, all methods can deal with the sitting condition without problems. The mean accuracy stays above 92% for all methods and the box plots fully remain at 100% with only a few outliers spread across the plot. The *Median Raw* performs with a mean accuracy of 96.5% and a mean error of 0.15 bpm better than all other methods, closely followed by the *Diff Model* and the *Diff Median* with a mean accuracy of 95.5% and 95%, and a mean error of 0.24 bpm and 0.49 bpm, respectively. In terms of signal quality, the *Diff Model* outperforms the other methods with a median PCC of 0.88 and a SNR of 35 dB. The *PCA*, *Mean Raw*, and *Median Raw* all comprise a median PCC of 0.81 and a SNR of 32 dB, while the remaining difference-based methods form the lower bound with a median PCC of 0.7 and a median SNR of about 30 dB, both with wide spread box plots. The *Diff Mean* and *Diff Median* most likely suffer from subtracting two noisy signals, hence increasing the overall noise level. The *Diff Model* with its built-in low-pass filter behaviour can reduce the noise sufficiently well and, moreover, also has a better signal quality than the non-difference-based methods.

Standing (Chest). Standing introduces slight motion artifacts that mainly are caused by small, unconscious swaying movements while keeping balance, but sometimes they also stem from the user relieving a leg, moving an arm by for instance taking the hands out of the pockets, or by changing its posture in general. The non-difference-based methods, i.e. the *PCA*, *Mean Raw*, and *Median Raw*, cannot compensate for these motion artifacts which leads to a mean accuracy of about 51% (median error in the same range) and a mean error of about 3.1 bpm to 3.8 bpm (median error 1.8 bpm to 2.1 bpm). The *Diff Mean* and *Diff Median* are able to subtract the motion components and thus are better in dealing with the standing condition. Their mean accuracy lies at about 93% (median 100%), and their mean error ranges from 0.4 bpm to 0.46 bpm. The *Diff Model* outperforms all other methods with a mean accuracy of 96% (median 100%) and a mean error of 0.22 bpm. In terms of signal quality, the non-difference-based methods show a median PCC of 0.21 and a median SNR of 17 dB. The *Diff Mean* and *Diff Median* have a median PCC of about 0.6 and a median SNR of about 29 dB. The *Diff Model* finally has a median PCC of 0.78 and a median SNR of 33 dB.

Occlusion (Chest). The drinking gestures cause even more body movements than standing alone and furthermore introduce regular self-occlusions through the arms and the cup held in the hands. The *PCA* and *Mean Raw* cannot compensate for any of these events and therefore only have a mean accuracy of about 21% (median 0%) with relatively large mean errors of 4.4 bpm to 4.7 bpm (median 3.4 bpm to 3.8 bpm). Their median PCC is at about 0.13 and their median SNR at about 13 to 14 dB. The *Median Raw*, to some extent, is more robust against deviating occlusion pixels and shows a mean accuracy of 45% (median 43%), a mean error of 1.9 bpm (median 1.8 bpm), a median PCC of 0.21, and a median SNR of 17.5 dB. As the median typically is more robust against outliers, median-based methods have a higher chance of not seeing an occlusion or of only suffering from it at a fraction of the time. The difference between using the mean or the median to extract the respiration signal in the presence of occlusion gets even more apparent when the body movement gets suppressed as by the *Diff Mean* and *Diff Median* methods. The *Diff Mean* performs worse than the *Median Raw*. It has a mean accuracy of 33% (median 14%), a mean error of 7.5 bpm (median 5.3 bpm), a median PCC of 0.13, and a median SNR of 13 db. The *Diff Median* on the other hand has a mean accuracy of 69% (median 93%), a mean error of 2.5 bpm (median 0.26 bpm), a median PCC of 0.28, and a median SNR of 20 dB. This finding strongly encourages the use of the median instead of the mean to estimate the breathing signal in the presence of occlusions. The *Diff Model* is able to detect and recover occluded body regions and therefore outperforms all other methods. It has a mean accuracy of 87% (median 100%), a mean error of 0.62 bpm (median 0.12 bpm), a median PCC of 0.61, and a SNR of 29 dB.

Dependency on Body Region. The overall influence of the body region already was explained in Section 5.5.3, so in this section the focus will lie on the interdependency of the condition and the body region. For this reason, the most important information about this interdependency is provided on a higher level, without going through all different performance values in detail. For reference, all performance measures can be found in Figure 5.16. At the abdomen and at the entire torso, but especially at the abdomen, all performance measures drop when compared to the chest. The

decrease in performance, however, is less marked during the sitting condition. Here, the *Median Raw* and the *Diff Model* can deal with the different body regions the best, while the *Diff Mean* shows the largest performance losses. During standing, the *PCA*, *Mean Raw*, and *Median Raw* show a weak performance on all regions. The *Diff Mean* and the *Diff Median*, while comprising a high performance at the chest, are strongly affected at the other body regions, mostly at the abdomen. The *Diff Model* can deal with the standing condition well when looking at the chest or the entire torso, but struggles at the abdomen. The difference-based method's decrease in performance during the standing condition is likely caused by the spatial distance of the reference region at the throat to the respective body region, like the abdomen. For the occlusion condition, only the *Diff Median* and the *Diff Model* are looked at. While the *Diff Median* gets severely affected at the torso and even more at the abdomen, the *Diff Model* can maintain an acceptable performance at the torso, but also struggles at the abdomen. During the occlusion condition, the methods do not only have to deal with the participants standing upright, as before, but also with a mug being held in one or both hands and being moved in front of the torso. Since the hand by most participants and most of the time was held in front of the abdomen and only occasionally was moved over the chest while performing a drinking gesture, the abdomen, but also the torso are prone to comprise a lot more motion artifacts than the chest.

Summary. While all methods are able to achieve high performance values during the sitting condition, a completely different picture is drawn at the other conditions. Standing introduces small motion artifacts which the *PCA*, *Mean Raw*, and *Median Raw* methods cannot compensate for. These motion artifacts thus interfere with the respiration signal and consequently their performance decreases significantly. The *Diff Mean*, *Diff Median*, and *Diff Model* are able to subtract the motion components from the signal and can, at least at the chest, maintain a comparably high performance as compared to the sitting condition. During the occlusion condition, the *PCA* and *Mean Raw*, as well as the *Diff Mean* again experience a significant drop in performance as compared to standing alone, while the *Median Raw* does not show such a high decrease in performance. As the median typically is more robust against outliers, the methods using the median have a higher chance of not seeing an occlusion or of only suffering from it at a fraction of the time. Consequently, the *Diff Median* is able to deal with the occlusions better than all methods mentioned above, but still it is heavily affected by the hand movements. The *Diff Model* on the other hand can handle the occlusions much better, but, to some extent, also experiences a drop in performance.

5.5.5 *The Influence of the Distance to the User*

There are two important factors that influence the breathing estimation when changing the distance of the user to the depth camera. First, with increasing distance the body region appears smaller on the image frame and fewer depth pixels are available for extracting the respiration signal. Secondly, the noise level of the depth camera's pixel readings increases with distance. Consequently, with increasing distance of the

user, a lower signal quality can be expected due to the decreasing amount of breathing related depth pixels available for averaging out the increasing noise. Another aspect is that in close proximity not all body joints may be visible, and on far distances the body joint estimation may not work due to too few body features being distinguishable on the smaller body appearance. For the Kinect SDK, the highest distance is at about 4 to 4.5 meters, and a minimum distance of 1 meter has been shown to be sufficiently far away during the experiments.

In this section, the influence of an increasing distance on the breathing estimation is evaluated. Figure 5.17 depicts the accuracies, errors, PCC, and SNR values at the chest of the different methods at distances ranging from 1 m to 4 m. The plots are separated into the three conditions sitting, standing, and occlusion. This ensures to not confuse the influence of the distance with a performance dependency on the condition and reveals the particular differences among conditions.

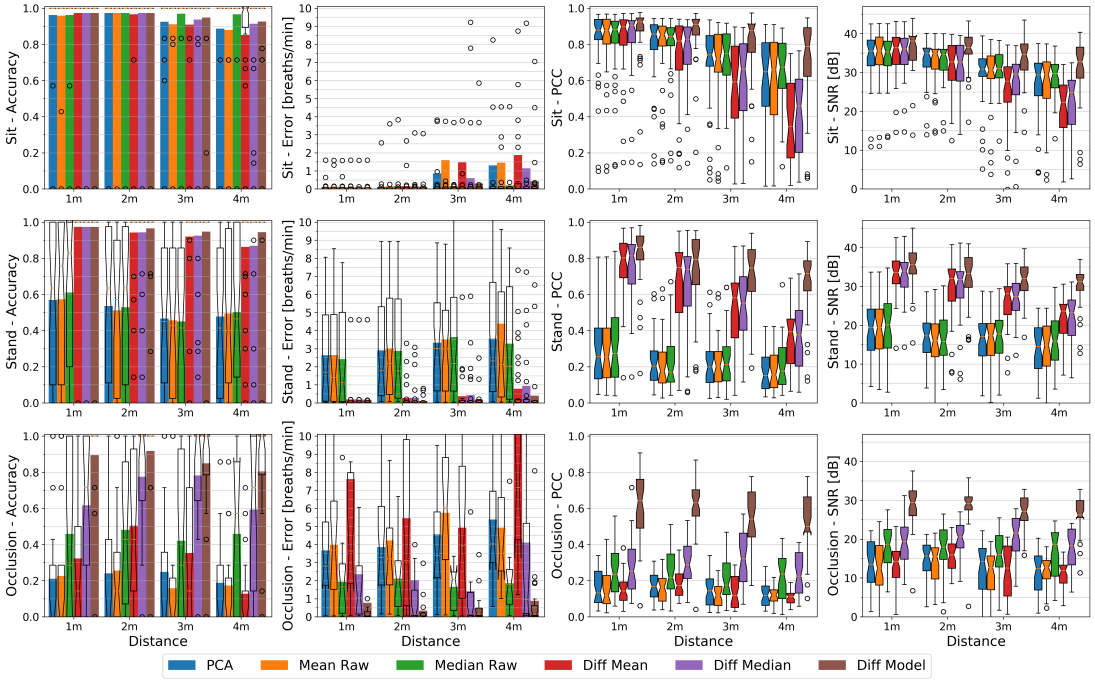

Figure 5.17: From left to right: The respiratory rate accuracy, errors in breath per minute, Pearson Correlation Coefficient w.r.t. the ground truth signal, and Signal-to-Noise Ratio of the different methods at the **chest**. Plots are separated from top to bottom by the sitting, standing, and occlusion condition and by the distance between 1 to 4 meters, with all respiratory rates (10 bpm and 15 bpm) combined. Accuracy and error metrics use a FFT window with a length of 48 seconds. The colored bars show the averages, while overlay box plots show median (middle parts) and whiskers marking data within 1.5 IQR. The breathing rate is detected slightly less accurately and the PCC and SNR values decrease when the user is further away from the camera. The *Diff Model* method remains robust for different distances and conditions.

Sitting. While sitting, all methods can maintain a median accuracy of 100% at all distances and except for the *Diff Mean* at 4 m, also all method's box plots fully remain at an accuracy of 100% with only a few more outliers at higher distances. The

mean accuracy of all methods likewise is highest at close distances, but decreases in varying amounts towards higher distances. At distances of 1 and 2 meters, all methods show a mean accuracy of about 96% to 97%. Their mean error at 1 m lies between 0.09 bpm and 0.12 bpm (medians at about 0.04 bpm) and increases at 2 m to about 0.12 bpm to 0.15 bpm (medians at about 0.05 bpm). From 3 m onwards, in terms of accuracy and error, a small but noticeable performance drop can be observed for most methods. The *Median Raw* hereby is minimally affected by the distance and is able to maintain a mean accuracy of about 97% and a maximum mean error of about 0.19 bpm at 4 m. The *PCA*, *Mean Raw*, and *Diff Mean* are affected the most. At 3 m they show a mean accuracy of about 91% to 92% with a mean error between 0.86 bpm for the *PCA*, up to 1.6 bpm for the *Mean Raw* (their median errors are at about 0.06 bpm). At 4 m, their performance values lower to a mean accuracy of about 85% for the *Diff Mean* and about 88% for the *PCA* and *Mean Raw*, with a mean error of 1.3 bpm for the *PCA* up to 1.9 bpm for the *Diff Mean* (all median errors at about 0.09 bpm). The *Diff Median* shows a mean accuracy and error of 94% and 0.6 bpm at 3 m, and 91% and 1.1 bpm at 4 m, and the *Diff Model* achieves 95% in accuracy and an error of 0.3 bpm at 3 m, and 93% and 0.4 bpm at 4 m, respectively.

In terms of signal quality, the PCC and SNR values also drop with increasing distance, but on a much larger scale than the accuracy, and with extending box plots towards higher distances. The median PCC of the *PCA*, *Mean Raw*, and *Median Raw* drop from a value of about 0.87 at 1 m to about 0.66 at 4 m, and their SNR drops from 35 db to about 29 db. The *Diff Mean* performs worst on higher distances with a median PCC ranging from 0.89 at 1 m to 0.34 at 4 m, and a median SNR from 36 db to 22 db. It is closely followed by the *Diff Median* with a median PCC range from 0.89 at 1 m to 0.45 at 4 m and a median SNR range from 36 db to 25 db. The *Diff Model* is least affected by the distance and spans from a median PCC of 0.92 at 1 m to 0.78 at 4 m and a median SNR from 38 db to 33 db.

Overall, the influence of the depth camera's increasing noise level at higher distances can best be observed in a seated position where the respiration signal is not disturbed by motion artifacts. When looking at the *PCA*, *Mean Raw*, or *Median Raw* methods, their PCC and SNR values get worse on higher distances whereas the *Diff Model* method with its inherent low-pass filtering remains more stable over all distances. The *Diff Mean* and *Diff Median* methods on the other hand decrease the most in signal quality due to computing the difference of two noisy signals, hence amplifying the overall noise. Both methods show the importance of low-pass filtering the depth values when using a difference-based approach for computing the respiration signal.

Standing. The standing condition introduces random body movements, for instance swaying while keeping balance, which have a dominating influence on all non-difference-based methods. Since the influence of the standing condition is not predictable and may vary in between different distances, the results of these methods have to be taken with caution. For this reason and because the non-difference-based *PCA*, *Mean Raw*, and *Median Raw* show similar performances on all measures, only their general trend is summarized, without listing them separately. Their mean accuracy is with about 59% highest at 1 m, drops to about 45% at 3 m, and interestingly increases again at 4 m to about 49%. This increase likely is caused by some partici-

pants moving less at 4 m, which also is supported by the difference-based methods that do not show such an increase. Their mean error increases from about 2.5 bpm (median 1.1 bpm to 1.7 bpm) at 1 m to about 3.5 bpm (median 2.2 bpm to 3.0 bpm) at 3 m. At 4 m, the *PCA* and *Median Raw* have a mean error of about 3.5 bpm and 3.3 bpm, and the *Mean Raw* of about 4.4 bpm (all medians at about 2.2 bpm). The *Diff Mean* and *Diff Median* show a similar mean accuracy on all distances that decreases from 97% at 1 m to 86% at 4 m. Their mean error increases from 0.18 bpm (median 0.06 bpm) at 1 m to 0.78 bpm for the *Diff Mean* and to 0.95 bpm for the *Diff Median* (all medians 0.08 bpm) at 4 m. The *Diff Model* also starts with a mean accuracy of 97% and a mean error of 0.18 bpm (median 0.06 bpm) at 1 m, but it only lowers to 95% and 0.4 bpm (median 0.07 bpm) at 4 m. The difference-based methods' accuracy box plots furthermore fully remain at 100% at all distances.

The median PCC and SNR values of the *PCA*, *Mean Raw*, and *Median Raw* methods lie between about 0.25 and 20 dB at 1 m and 0.16 and 15 dB at 4 m, all indicating a poor signal quality. The *Diff Mean* and *Diff Median* start with a median PCC and SNR of 0.81 and 34 dB at 1 m and drop to about 0.38 and 23 dB at 4 m, which is a similar trend as for the sitting condition. With increasing distance, the *Diff Model* also loses in signal quality, but with a median PCC and SNR between 0.86 and 36 dB at 1 m, and 0.72 and 31 dB at 4 m, it performs significantly better than the other methods. Being able to maintain a better signal quality especially at higher distances also explains its higher accuracy as compared to the other difference-based methods.

Occlusion. With the introduction of self-occlusion events, it is barely possible to draw any conclusions about the influence of the distance on methods that are not able to deal with those. The reason is that random amounts, extents, and times of the occlusions on top of random movements caused by staying enter the breathing signal in an unpredictable way. Recordings at higher distances might comprise less motion artifacts and thus are likely to yield better performance values than recordings from close distances, or vice versa. These random signal distortions therefore are likely to shadow any effects of the distance when not counteracted.

The *PCA* and *Mean Raw* have a mean accuracy below 25% at all distances and the *Median Raw* shows values between 42% and 48% randomly distributed between 1 m and 4 m. The *Diff Mean* has a maximum mean accuracy of about 51% at 2 m which to both sides degrades to below 35% down to about 13% at 4 m. Except for the *Median Raw*, all these methods have a mean error above 3.7 bpm, a median PCC below 0.17, and a median SNR below 16 dB. The *Median Raw* performs better than above methods and shows a mean error of between 1.7 bpm (3 m) to 2.1 bpm (2 m), a median PCC of about 0.21, and a median SNR of about 18 dB across all distances. The *Diff Median*, as already described in Section 5.5.4, can deal with the occlusion scenario much better. Starting with a mean accuracy of 62% (median 71%) at 1 m, it achieves up to 78% (median 100%) at distances from 2 m to 3 m, and falls down to 60% (median 71%) at 4 m. Its mean error decreases from 2.4 bpm (median 1.0 bpm) at 1 m to 1.4 bpm (median 0.15 bpm) at 3 m and increases to 4.1 bpm (median 0.93 bpm) at 4 m. Its median PCC and SNR likewise increase in between 1 m to 3 m from 0.25 to 0.32 or from 19 dB to 22 dB and have a reduced value of 0.22 or 18 dB at 4 m. The *Diff Model*, in contrast to the other methods, is able to detect and recover occluded areas. It has a mean accuracy of about 90% at 1 m, 92% at 2 m and drops to about 80% at 4 m, with

all median accuracies at 100%. Its mean error at 1 m is 0.77 bpm (median 0.1 bpm) and gradually increases from 0.33 bpm (median 0.1 bpm) at 2 m to 0.88 bpm (median 0.15 bpm) at 4 m. Its median PCC and SNR drop from 0.64 and 29 dB at 1 m to 0.5 and 26 dB at 4 m. All methods show a decreased performance at 1 m as compared to a distance of 2 m, which to some extent is likely to be caused by the randomness of the occlusion gestures. Another explanation may be that the occluding hand and mug block at a closer distance more infrared rays emitted by the depth sensor and cast a shadow on nearby pixels. An additional reflection of the emitted infrared rays from the mug towards the body furthermore influences a certain non-occluded area on the body surface.

Summary. The optimal distance to measure the breathing signal has been shown to be in the range from 1 m to 2 m, with a tendency towards 2 m in case of occlusion events. A greater distance hereby mainly affects the signal quality as can best be observed for all methods when looking at the PCC and SNR values of the sitting condition. While sitting, the respiration signal is not disturbed by motion artifacts and thus only competes against the increasing noise level of the depth camera at higher distances. The reduced signal quality due to the increasing noise then in return has an effect on the accuracy and error rate. The type of the condition, however, has a much stronger influence than the distance. All methods that are not designed to deal with motion artifacts or occlusion show on all distances a significantly reduced performance by means of accuracy, error, and signal quality (also see Section 5.5.4). Due to the randomness of these signal distortions, for these methods it furthermore is barely possible to draw any conclusions about the influence of the distance, while the methods that can deal with the respective condition show a similar trend as observed for the sitting condition. In the occlusion scenario, a distance of 1 m has shown to be less optimal as compared to a distance of 2 m, which is assumed to be due to increased shadowing and reflection effects caused by the occluding hand and mug from the depth camera's infrared emitter upon the body surface.

5.5.6 *The Influence of the Respiratory Rate*

The sitting and standing sessions were recorded at two different respiratory rates of 10 breaths per minute (0.17 Hz) and 15 breaths per minute (0.25 Hz), both obtained from the paced breathing setup. In this section, the influence of the respiratory rate on the different methods' performances is assessed and quantized using the accuracy, error, PCC, and SNR. The performance values are taken from the chest region, include all distances, and are separated into the conditions sitting and standing. The results for the two different respiratory rates are depicted in Figure 5.18.

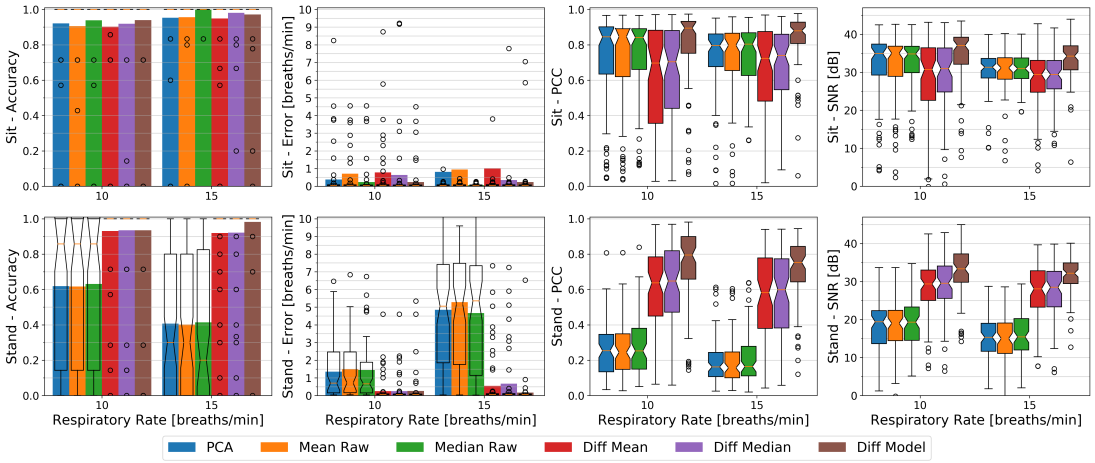

Figure 5.18: From left to right: The respiratory rate accuracy, errors in breath per minute, Pearson Correlation Coefficient w.r.t. the ground truth signal, and Signal-to-Noise Ratio of the different methods at the **chest**. Plots are separated from top to bottom by the sitting and standing condition and by the respiratory rate of 10 bpm or 15 bpm, with all distances (1 to 4 meters) combined. Accuracy and error metrics use a FFT window with a length of 48 seconds. The colored bars show the averages, while overlay box plots show median (middle parts) and whiskers marking data within 1.5 IQR. The higher respiratory rate show a slightly better accuracy over all methods when sedentary, but a lower signal quality (PCC and SNR values) on both conditions.

Sitting. While sitting, and at a respiratory rate of 10 bpm, the *Median Raw* and the *Diff Model* achieve the highest performance with a mean accuracy of about 94% and a mean error of 0.25 bpm. They are followed by the *PCA* and *Diff Median* with a mean accuracy of 92% each, and a mean error of 0.39 bpm and 0.63 bpm respectively. The *Mean Raw* and *Diff Mean* show the lowest performance with a mean accuracy of about 91% and 90%, and a mean error of about 0.72 bpm and 0.78 bpm, respectively. At 15 bpm, the highest mean accuracy and lowest mean error of about 100% and 0.06 bpm is achieved by the *Median Raw*, closely followed by the *Diff Median* with 98% and 0.35 bpm, and the *Diff Model* with 97% and 0.24 bpm. The remaining methods have a mean accuracy of about 95%, with the mean error of the *PCA* being at 0.82 bpm, of the *Mean Raw* at 0.95 bpm, and of the *Diff Mean* being at 1.0 bpm. Furthermore, at both respiratory rates, all methods' accuracy box plots fully remain at 100% and all methods show a median error of about 0.05 bpm. At 10 bpm, however, much more outliers can be observed in the error plot, with most of them falling in the range of up to an error of about 4.5 bpm, hence decreasing the respective methods' mean accuracy at the 10 bpm breathing rate.

In terms of signal quality, the *PCA*, *Mean Raw*, and *Median Raw* have a median PCC and SNR of about 0.85 and 35 dB at 10 bpm, and 0.8 and 30 dB at 15 bpm. The *Diff Mean* and *Diff Median* show a lower signal quality with values of 0.7 and 31 dB at 10 bpm, and 0.73 and 29 dB at 15 bpm. The *Diff Model* has on both respiratory rates the highest median PCC and SNR with values of 0.89 and 37 dB at 10 bpm, and 0.88 and 34 dB at 15 bpm.

Standing. While standing, the non-difference-based methods, as explained in Section 5.5.4, are heavily influenced by that condition and show a low performance, but a strong influence of the respiratory rate can be observed. The mean accuracy of the *PCA*, *Mean Raw*, and *Median Raw* at 10 bpm is with about 62% (median 86%) much higher than at 15 bpm where it only is at about 41% (median 20% for *Median Raw*, 30% others). The mean error likewise is for these methods with about 1.4 bpm to 1.5 bpm (median 0.7 bpm) lower at 10 bpm than at 15 bpm where it is above 4.7 bpm (median above 5.0 bpm). Likewise, their median PCC and SNR at 10 bpm indicate with values of 0.25 and 19 dB a better signal quality than at 15 bpm, which in contrast shows lower PCC and SNR values of 0.15 and 15 dB.

The difference-based methods are not or only barely affected by the standing condition. The *Diff Mean*, *Diff Median*, and *Diff Model* methods show a mean accuracy of about 93% and a mean error of 0.25 bpm at a respiratory rate of 10 bpm. At 15 bpm, the *Diff Mean* and *Diff Median* have a slightly lower mean accuracy of about 92% and a higher mean error of 0.55 bpm and 0.67 bpm, respectively. The *Diff Model* on the other hand achieves at 15 bpm a higher mean accuracy of 98% and a lower mean error of 0.17 bpm. On both respiratory rates, the difference-based methods furthermore show a median error of 0.06 bpm and have their accuracy box plots being fully at 100%. The median PCC and SNR values of the *Diff Mean* and *Diff Median* are at about 0.64 and 29 dB for the 10 bpm and at about 0.59 and 28 dB for the 15 bpm rate. The *Diff Model* has the highest PCC and SNR values of 0.8 and 33 dB at 10 bpm, and 0.75 and 32 dB at 15 bpm.

Summary. All methods appear to have a lower signal quality at 15 bpm as compared to 10 bpm as indicated by both, the Pearson Correlation Coefficient and the Signal-to-Noise Ratio. All methods' mean accuracy values on the other hand are higher at 15 bpm during the sitting condition and for the *Diff Model* during the standing condition. A likely reason for this is that more signal periods fall within the 48s FFT window at 15 bpm than at 10 bpm, making the 15 bpm signal component stronger and easier to detect in frequency domain, at least during the sitting condition with weak frequency components stemming from motion artifacts. Since the differences in accuracy are not that big, they might, however, also be caused by one or a few users. For the other cases, it can be argued that the higher respiration frequency interferes stronger with other body movement and thus can not be detected that easily, but it also is likely that the relatively relaxed low respiration frequency of 10 bpm (0.17 Hz) did not introduce as many motion artifacts as the faster one, or that it was easier to maintain during the recording.

5.5.7 *The Influence of the Gender*

The sex or gender can be considered an important distinguishing feature between different users. Male and female persons do not only differ in body shape, but also otherwise typically show distinct differences in their appearance. Most notably are different clothing styles and the tendency of female persons to have longer hair than their male counterparts, but also the overall body posture tends to be different [127]. All these characteristics influence the torso appearance on the depth data and thus

can be assumed to also have an influence on the respiration estimation. Note that the term gender here is used over the term sex since clothing styles, length of hair, etc. are not dependent on the biological sex, but on the social norms and roles associated with the gender. Since social norms can change over time and from culture to culture, here it is referred to typical gender norms and roles found in Germany at the time of writing.

To assess gender-specific differences on all methods' performance, the participants are split into a male and a female group, each containing 12 male or 7 female participants, respectively. Figure 5.19 depicts the accuracy, error, PCC, and SNR of the different methods for both groups, again divided into the three conditions sitting, standing, and occlusion.

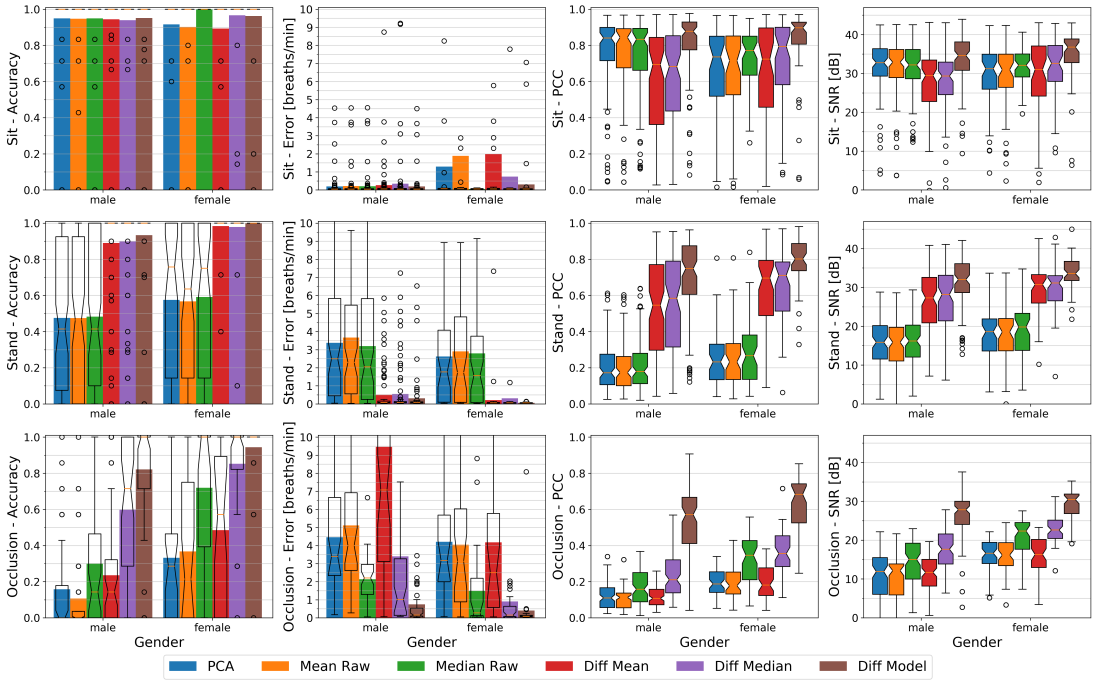

Figure 5.19: From left to right: The respiratory rate accuracy, errors in breath per minute, Pearson Correlation Coefficient w.r.t. the ground truth signal, and Signal-to-Noise Ratio of the different methods at the **chest**. Plots are separated from top to bottom by the sitting, standing, and occlusion condition and by the gender of the participants, with all distances (1 to 4 meters) and all respiratory rates (10 bpm and 15 bpm) combined. Accuracy and error metrics use a FFT window with a length of 48 seconds. The colored bars show the averages, while overlay box plots show median (middle parts) and whiskers marking data within 1.5 IQR. Except for the sitting condition, all considered methods perform significantly better on the female participants, mostly due to unbalanced group sizes (12 male, 7 female) and highly variable results for the male participants (user 9, male, had problems in adhering to the breathing visualization, also see Section 5.5.8).

Sitting. While sitting, the male users show a mean accuracy of about 95% (box plots fully at 100%) and a mean error of 0.19 bpm to 0.34 bpm (median 0.05 bpm) for all methods. In the female group, the *PCA*, *Mean Raw*, and *Diff Mean* achieve the lowest mean accuracy of about 90% to 92% and mean errors between 1.29 bpm to

2.0 bpm (medians at 0.07 bpm), whereas the *Median Raw* achieves with values of 100% and 0.06 bpm (median 0.05 bpm) the highest performance. The *Diff Median* and *Diff Model* show a mean accuracy of about 96% and a mean error of 0.74 bpm (median 0.06 bpm) and 0.31 bpm (median 0.05 bpm), respectively. The accuracy box plots for all methods, like on the male group, nonetheless fully remain at 100%. When zooming out of the female's error plot, a few outliers can be seen, probably from a single person, with values of up to 30 bpm that, in combination with their smaller group size, cause the mean error of the *PCA*, *Mean Raw*, *Diff Mean*, *Diff Median* to be much higher as compared to the male group.

In terms of signal quality, the PCC and SNR of the non-difference-based methods are with values of 0.84 and 33 dB higher for the male group than compared to the female group with PCC values of 0.71 to 0.77 and a SNR of about 31 dB to 32 dB. The difference-based methods on the other hand show a lower signal quality on the male group. Here, the PCC of the *Diff Mean* and the *Diff Median* have a value of 0.68 and a SNR of 29 dB as compared to a PCC of 0.72 or 0.8 and a SNR of 31 dB or 32 dB. The *Diff Model* achieves the highest signal quality, with PCC and SNR values of 0.88 and 35 dB for the male, and 0.9 and 37 dB for the female group.

Standing. The standing condition, as already mentioned in Section 5.5.4, is challenging for the non-difference-based methods. All of them show a mean accuracy of about 48% (medians slightly below) and mean errors between 3.2 bpm to 3.7 bpm (median 2 bpm to 2.5 bpm) for the male, and about 58% (median 63% to 74%) and 2.6 bpm to 2.9 bpm (median 1.5 bpm to 1.7 bpm) for the female users. Their median PCC and SNR values are slightly higher for the female users, but are all below 0.27 and 20 dB. In contrast to that, all difference-based methods have accuracy box plots that fully remain at 100% and median errors below 0.07 bpm. The *Diff Mean* and *Diff Median* methods show similar values per group. Their mean accuracy and mean error values lie at about 90% and at 0.51 bpm to 0.55 bpm (medians 0.07 bpm) for the male, and at about 98% and at 0.21 bpm to 0.30 bpm (medians 0.05 bpm) for the female group. The *Diff Model* has a mean accuracy of 93% and a mean error of 0.31 bpm for the male, and 100% and 0.06 bpm for the female group.

In terms of signal quality, the *Diff Mean* and the *Diff Median* show PCC and SNR values below 0.58 and of about 28 dB for the male, and below 0.7 and about 31 dB for the female users. The *Diff Model* has a median PCC and SNR of 0.74 and 32 dB for the male, and 0.8 and 33 dB for the female group.

Occlusion. As a general trend during the occlusion condition, it can be seen that all methods show a higher performance on the female users than on the male group. The *PCA*, *Mean Raw*, and *Diff Mean* methods hereby perform worse than the other methods and can be considered to be more susceptible to occlusions than the *Diff Model* or both median based methods. Due to the randomness of the occlusion gestures and considering the different group sizes (12 male, 7 female), it furthermore is hard to derive an influence of the gender for all methods that are susceptible to occlusion events. For this reason, at this point no conclusions are drawn about the influence of gender on the *PCA*, *Mean Raw*, and *Diff Mean* methods. Also the median based methods have to be taken with care, but since they show reasonable results and big differences on both groups, both methods are examined more closely. The

Median Raw jumps from a mean accuracy of 30% (median 14%) and a mean error of 2.1 bpm (median 2.2 bpm) on the male group to 72% (median 100%) and 1.5 bpm (median 0.26 bpm) on the female group. Its accuracy and error for the female group are even better than its performance during the standing condition, but the occlusion condition also only was recorded at a respiratory rate of 10 bpm, which according to results from Section 5.5.6 is known to yield higher performance values. The *Diff Median* similarly performs better on the female group, where it shows a mean accuracy of 85% (median 100%) and a mean error of 0.9 bpm (median 0.15 bpm) as compared to values of 60% (median 71%) and 3.4 bpm (median 1.0 bpm) on the male group. In terms of accuracy, the *Diff Median* performs on the female group even better than all other methods on the male group. The *Diff Model* achieves for the female participants with a mean accuracy of 94% (median 100%) and a mean error of 0.41 bpm (median 0.08 bpm) the highest performance among all methods and groups, whereas for the male participants it yields values of 82% and 0.75 bpm (median 0.15 bpm).

When looking at the signal quality measures, it can be observed that for the female users also a higher signal quality can be obtained by all methods. On the female group, the *Median Raw*'s PCC and SNR median values are at 0.34 and 22 dB, and on the male group they are at 0.15 and 15 dB. Similarly, the *Diff Median* has median PCC and SNR values of 0.35 and 22 dB on the female, and 0.21 and 17 dB on the male group. Despite the relatively good accuracy and error performance, these PCC and SNR values suggest a rather low signal quality. In terms of signal quality, the *Diff Model* stands out from the rest. It has median PCC and SNR values of 0.68 and 30 dB for the female, and 0.56 and 28 dB for the male users.

Summary. On first sight, it seems like all methods work better on the female participants than on the male ones, especially in the standing and the occlusion scenarios. The male group, however, is with 12 participants almost twice as big as the female group. With only 7 female participants in the dataset, it is consequently hard to pinpoint whether the user's gender could play a role in the performance of breathing rate estimation. Due to the relatively small and unbalanced group sizes, it is likely that the performance is biased towards the female group. Also, in the used dataset, in contrast to some male participants, all females had clothing that did not cover the throat. So at least in this dataset the gender-specific differences might not be caused by the gender itself but by gender specific clothing styles (also see Section 5.5.8). Furthermore, a single male user was found that had difficulties in adhering to the paced breathing setup and, due to the small group size, lowers the overall performance of the whole male group. The influence of single users and specific properties like their clothing styles will be elaborated in the next section.

5.5.8 *The Influence of the User*

In this section, the influence of the single users on the performance of the various respiration estimation methods is assessed. The results of this section are meant to give some context to the different evaluation outcomes from previous sections and should not be seen as definitive results, but rather as an indicator for future research questions.

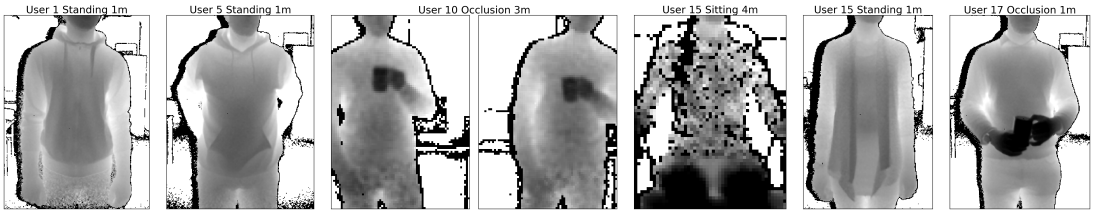

Figure 5.20: Example images from the recordings of (from left to right) user 1, user 5, user 10, user 15, and user 17. Their clothing is particularly challenging for different types of methods. Difference-based methods have difficulties when the throat is partially covered by a collar (user 1, user 5, and user 17), resulting in a poor motion reference signal obtained from the throat region. Clothing with bad infrared reflection properties (user 15, left) and garments loosely hanging from the chest or the shoulders (user 15, right) or similar clothing that causes many surface deformations over time, like ribbons reaching to the chest area (users 1 and 5), are likely to interfere with the respiration signal obtained from all methods. Also hair covering the body surface (user 15, left) or users moving a lot or bending to either side (user 10) are likely to cause signal distortions.

There is a whole set of user-specific parameters that may directly or indirectly influence the measurements. These include size and weight, state of health, age, gender, clothing, or even long hair reaching to the chest area, but also the preferred breathing rhythm and style, e.g. abdominal breathing, or simply the ability to stand still for a while. Since the dataset focuses on having a high user variance in order to achieve meaningful results in above parameter evaluations, the participants were not explicitly categorized by these parameters. Furthermore, a systematic evaluation by for instance asking the users to wear a specific set of different clothing styles was not pursued. Consequently, each participant shows a rather unique subset of user-specific parameters. Due to the big parameter space and the limited number of participants, it therefore is difficult to draw final conclusions about user-specific influences, as mentioned at the beginning of this section.

An attempt to nevertheless gain an insight into user specific parameters thus is as follows: If a user can be identified that, regardless of the method used, performs worse than other users, this user may exhibit a specific reason for why he or she influences the respiration estimation. Furthermore, previous evaluations hide the contributions of single participants to the average values and box plots. By inspecting the data on a per user basis, more detailed information about the composition of these plots can be obtained, like if a lower performance is caused by all participants similarly, or if one or a few participants with exceptionally low performance values cause a significant decrease on the averages.

5.5.8.1 *Accuracy and Error of Users*

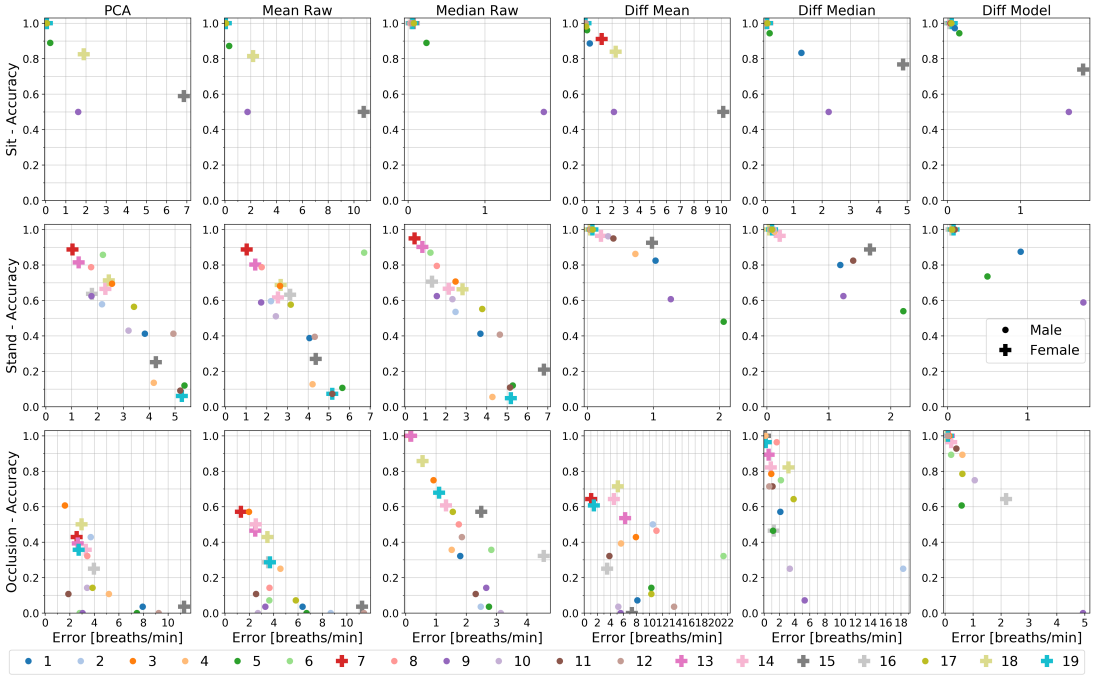

Figure 5.21: The mean accuracy (Y-axis) over the mean error (X-axis) for all individual users, captured from the **chest** using a FFT window with a length of 48 seconds. Accuracy and errors are averaged over all distances (1 to 4 meters) and all respiratory rates (10 bpm and 15 bpm) for the methods (from left to right): *PCA*, *Mean Raw*, *Median Raw*, *Diff Mean*, *Diff Median*, and *Diff Model*, each divided into the conditions sitting, standing, and occlusion. Each user is associated with a unique color where circles mark male and plus signs mark female users. Please note the different scales on the X-axis (errors). Ideally, all users are located at the top left corner, indicating a high accuracy and low error. This is for most participants achieved by all methods during the sitting condition. On other conditions, the *Diff Model* method achieves the best performance. The accuracy and errors can be seen to vary widely for certain participants, with especially user 9 standing out on all conditions and methods due to poorly adhering to the breathing visualization.

Figure 5.21 depicts for each participant the accuracy against the error, split up into the three conditions and averaged over all distances and respiratory rates. The single users are color-coded and marked with a dot for male, and with a plus for female users. Ideally, all user markings are at the upper left corner of the plots, where they indicate a high accuracy and low error on average.

Sitting. While sitting, almost all users show for all methods on average a high accuracy and a low error close to 100% or 0 bpm, respectively. Most notably, users 9 and 15 stand out on all methods. User 9 hereby shows a constant accuracy of 50% and an error of about 1.5 to 2 bpm for all methods. After inspecting this user's depth videos, user 9 was found to not or to only poorly maintain the respiratory rate given by the breathing visualization on almost all recordings, as will also be seen on the other

conditions. User 15 shows varying accuracy values between 50% and about 80% and high errors of up to 10 bpm on all methods, except for the *Median Raw* where she achieves 100% and close to 0 bpm error. The decreased performance is caused by poor infrared reflection properties of her clothing and her long hair partly covering her chest at distances from 3 m upwards as depicted in Figure 5.20. The *Median Raw* is able to achieve a high performance due to the median being more robust against this kind of noise where less than half the pixels are affected.

Standing. While standing, the *PCA*, *Mean Raw*, and *Median Raw* have difficulties in estimating the respiration as discussed in Section 5.5.4. The additional motion artifacts cause a wide, diagonally distributed spread of the user's performances towards low accuracy and high errors. The focus therefore lies on the difference-based methods where most participants again show a high accuracy of almost 100% and low error close to 0 bpm, especially for the *Diff Model*. On this condition, most notably users 1, 5, 9, and 15 stand out from the rest. User 9 again did only poorly maintain the given respiratory rate, and user 15 was recorded on a different day with a different, but nevertheless challenging dress: A cardigan with an open front hanging loosely from the shoulders as depicted in Figure 5.20 (2nd from the right). The *Diff Model* can compensate for the garment's movement during breathing due to its capability of detecting and recovering such occlusion events. Users 1 and 5 both have in common that they are wearing a hooded sweater that partly covers the throat region (see Figure 5.20) where the motion reference signal is extracted from. Especially at higher distances, this region resolves to only a few pixels and a moving collar (due to chest expansion while breathing) is likely to interfere.

Occlusion. During the occlusion condition, except for the *Diff Model*, all methods show a wide spread of the users' accuracy and error averages, with the *PCA*, *Mean Raw*, and *Diff Mean* only achieving a maximum accuracy of about 60% or 75% on a few participants. The *Diff Median* performs significantly better with most users above 70% and up to 100%. Compared to the *Diff Model* that is able to shadow occlusion events, it however can not compete, so the focus will be on the *Diff Model* only. Here, users 5, 9, 10, 16, and 17 deviate most significantly from the other users, which in contrast to those all lie in the range from 90% to 100% and below 1 bpm. User 5 again is likely to only achieve an average accuracy of about 60% due to the hooded pullover with the collar covering the throat region, and user 9 again had difficulties to adhere to the breathing visualization. User 10, in contrast to other users, occasionally shows strong movements to either of both sides while relieving a leg. These movements directly affect the respiration signal and lower the performance, most likely due to window misalignment caused by bending the upper body to the side or by the quickness of the leaning movement. For user 16, it was found that the drinking gestures were not fully executed with the cup often remaining for longer time periods in front or close to the throat region, which decreases the performance at distances of 3 and 4 meters where this region is only a few pixels wide. User 17 is wearing a shirt with a collar that also partly covers the throat (see Figure 5.20, right) and it was found that the decreased performance solely stems from the distance at 4 meters, again likely due to the lower resolution at higher distances with only a few pixels available to sample the motion signal from the throat region.

5.5.8.2 *Accuracy Distribution of Users*

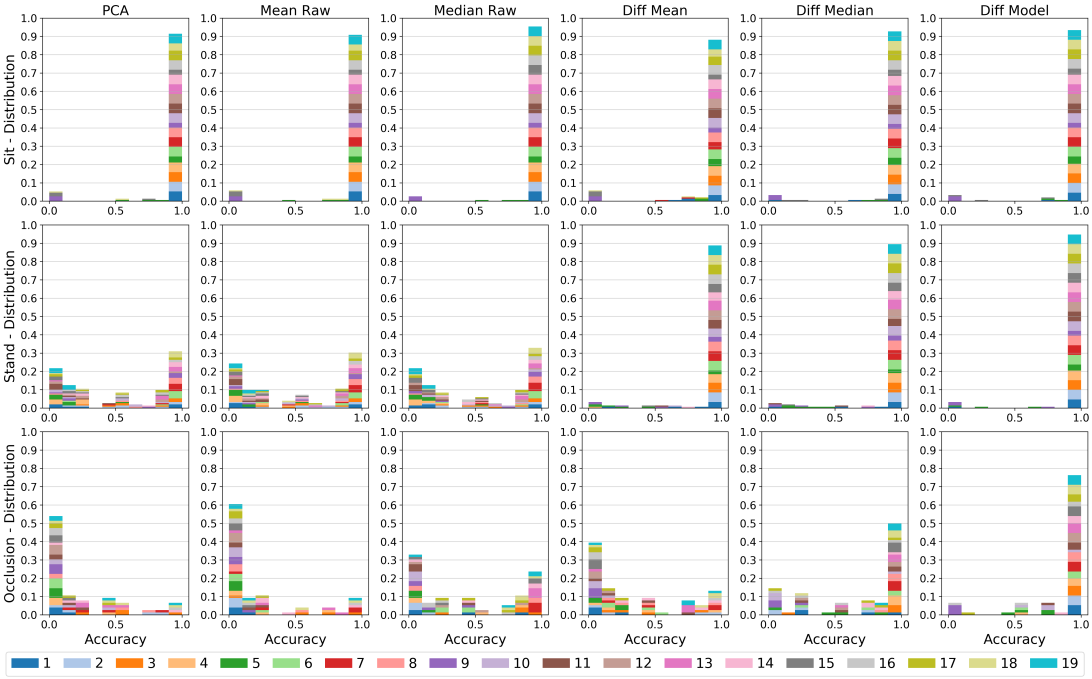

Figure 5.22: Histogram of the accuracy distribution of the single participants, with the breathing signal captured from the **chest** and using a FFT window with a length of 48 seconds. Histograms are divided into the different methods (from left to right) *PCA*, *Mean Raw*, *Median Raw*, *Diff Mean*, *Diff Median*, and *Diff Model*, and into the conditions (top to bottom) sitting, standing, and occlusion. The X-axis indicates the accuracy value divided into 10 bins and the Y-axis indicates the percentage of recordings that achieve this accuracy. The contributions of the individual participants are color-coded and stacked on top of each other. Ideally, all individual accuracy contributions are located in the rightmost bin, indicating that the respiratory rate can correctly be estimated on all recordings. A high bar in the 90-100% accuracy bin can be achieved by all methods during the sitting condition and by the difference-based methods during the standing condition. During the occlusion condition, the 0-10% accuracy bin is dominant for all methods, except for the *Diff Median* and the *Diff Model*, which have 50% or 76% of the recordings in the 90-100% bin.

Figure 5.22 depicts a histogram of the accuracy distribution of all participants. The whole accuracy range of the single recordings hereby is divided into 10 bins (or 10% steps) and drawn on the X-axis, while the Y-axis indicates the percentage of all recordings that achieve an accuracy that falls into the respective bin. Each user's contribution to a specific bin or accuracy range is visualized by color-coding the participants as before. The contributions of the single participants to a specific accuracy range are stacked on top of each other such that each participant's individual contribution as well as the overall amount of recordings with an accuracy that falls inside that bin is visualized.

Sitting. During the sitting condition, the majority of the users are within the 90% to 100% accuracy range, regardless of the method. Only user 9 constantly shows a significant part in the 0% to 10% range on all methods due to him not properly adhering to the breathing visualization. The same applies to the standing and occlusion conditions. User 15 notably also shows up with a significant part in the same accuracy range, but only for the *PCA*, *Mean Raw*, and *Diff Mean* methods.

Standing. During the standing condition, the *PCA*, *Mean Raw*, and *Median Raw* methods draw a completely different picture: Only about 30% of the recordings reach the 90% to 100% accuracy range and already about 22% of the recordings are in the 0% to 10% range. Notably, these methods work well for always the same users, even if to varying degrees. To some extent, thus a clear user dependency can be stated here, likely due to the respective participants being able to stand still for a while or due to a different breathing style. The difference-based methods do show a similar accuracy distribution like during the sitting condition, but with a few more participants partly showing up in lower accuracy ranges. With the *Diff Mean* and *Diff Median* methods, about 89% of the recordings, and with the *Diff Model*, about 95% of the recordings reach the 90% to 100% range. User 9 again constantly appears with a significant part at the lower end of the accuracy ranges, regardless of the method used.

Occlusion. During the occlusion condition, the 0% to 10% accuracy range becomes the dominant region for the *PCA*, the *Mean Raw*, the *Median Raw*, and the *Diff Mean* methods. Only the *Median Raw* shows with about 25% of the recordings a significantly higher peak in the 90% to 100% bin, where users 7 and 13 stand out as strongest contributors to that range. The *Diff Median* achieves to put about 50% of the recordings and most of the users into the 90% to 100% accuracy range. The other ranges do not show a significant peak and less than 15% of the recordings fall into any one of the remaining ranges. The *Diff Model* outperforms all other methods, with about 76% of the recordings and almost all users being in the 90% to 100% range. Moreover, the 0% to 10% accuracy bin almost completely is covered by user 9, the user that did not adhere to the breathing visualization, and the next peak with a comparably high impact (about 6% to 7% of the recordings) is in the 50% to 60% accuracy bin.

5.5.8.3 *User Summary*

In conclusion, it could be seen that in the cases where the method in use is suited for the condition, like the difference-based methods during standing or the *Diff Model* in the occlusion scenario, a decreased performance on any method mostly stems from a few individual users. The other way round, during the standing condition it could be seen that methods that are not suited for that condition show good results for a few, but always the same users, regardless of the method used. These findings indicate a dependency of depth-based respiration estimation on user-specific parameters.

Most notably, user 9 stands out from the rest on all methods and all conditions. This user did not or did only poorly maintain the respiratory rate given by the breathing visualization and thus decreases the performance measures of all methods by a certain amount, especially when comparing the male to the female group. This user

should be excluded from the dataset since this user's data does not reflect the correct respiratory rate as required from the study design, but nevertheless was kept in order to account for more realistic scenarios where users do not behave as expected. For other users, mainly the clothing style was found to be the most likely reason for a decreased performance. Examples are clothing with poor IR reflection properties (see Figure 5.20, fifth from the left), or loose garment or ribbons hanging from the chest or shoulders (see Figure 5.20, first, second, and sixth from the left). A collar that partly covers the throat region is likely to affect the difference-based methods because it tends to move during breathing and interferes with the motion reference signal extracted from the throat region (see Figure 5.20, first, second, and seventh from the left). Apart from these observations, there might also be more clothing-related factors like strong surface deformations that affect the estimation of the respiratory rate. For a full understanding of the influence of clothing, however, a separate, systematic study where the same participants are recorded with a set of different cloths needs to be conducted. Other user-specific influences that possibly affect the breathing estimation are long hair reaching to the chest (see Figure 5.20, fifth from the left) and movements such as changing the leaning angle to either side (see Figure 5.20, second and third from the left), or the ability to stand still for a while in general.

A detailed and systematic evaluation of user-dependent influences, for instance evaluating different clothing styles, is required in the future to fully understand the particular influences. This will enable the implementation of an adaptive method that for instance considers multiple body regions and rejects strongly influenced parts.

5.6 APPLICATIONS OF DEPTH-BASED RESPIRATION ESTIMATION

In this section, an outlook on applications of depth-based respiration estimation methods will be given with two examples based on two small-scale user studies. The examples are the utilization of remotely monitored respiration signals in activity recognition and their usage for medical applications, i.e. e-health and telemedicine.

5.6.1 *Remote Respiration Estimation as a Modality for Activity Recognition*

As many RGB-D approaches for activity recognition rely on extracting body pose (and sequences thereof), respiration features could also be seen as an additional modality to specify certain activities. Examples could be the quality of breathing for weightlifting exercises, the regularity of breathing for meditation, or the breathing speed reflecting affect or drowsiness while driving car or watching TV. Using respiration as an additional modality is likely to enhance the field of activity recognition in certain tasks, as it provides feedback on a user's condition, such as state of health, effort, affect, or drowsiness, and enables direct bio-feedback for these. Furthermore, it does not bring any additional hardware requirements in depth-based activity recognition scenarios while being non-intrusive and computationally efficient. It this section, it will be shown how the obtained respiration signal from the proposed method can be used to generate breathing-related features that characterize and separate several breathing-specific activities that would otherwise, for instance by observing the user's body pose, be hard to detect.

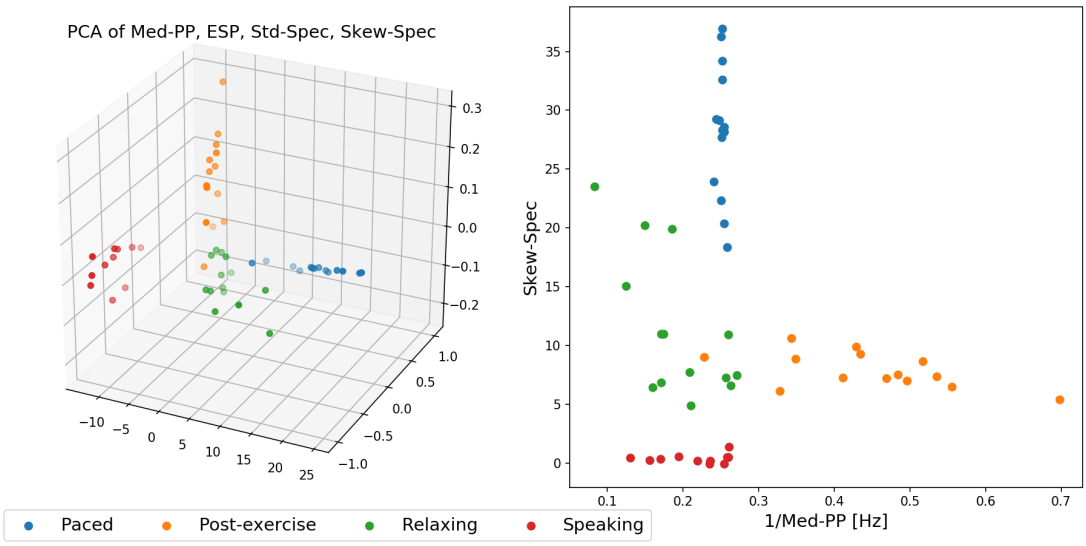

Figure 5.23: The different activities: paced-breathing meditating, post exercise recovering, relaxing, and speaking (reading aloud), transformed to feature space. **Left**: The features Med-PP, ESP, STD-Spec, and Skew-Spec reduced by one dimension by a principal component analysis. **Right**: The inverse of the feature Med-PP (to express it in Hz) plotted against Skew-Spec.

For the experiment, the dataset from the validation study is used, in which 14 participants performed the activities paced-breathing meditating, relaxing, and post exercise recovering (also see Section 5.3.2). In addition to these, a reading activity was added, in which 11 participants were asked to read aloud the same text across all study participants for several minutes. The reading activity was recorded under the same conditions as the three other activities (see Section 5.3.2). All participants were in a standing posture during all these activities, making it hard to use body posture or body joint sequences to be used to distinguish between these activities.

Several features that work particularly well to discriminate between the different activities were identified as follows:

- Standard deviation *Std-Spec*, skew *Skew-Spec*, and kurtosis *Kurt-Spec* of frequency spectrum amplitudes.
- Signal-to-Noise ratio *SNR*.
- Median of the time deltas between peaks in the signal *Med-PP*.
- Spectral entropy of the signal obtained in frequency domain *ESP*.
- Standard deviation of the first order time derivative of the signal *Std-Deriv*.

Figure 5.23 depicts two plots of the data expressed in feature space using the more promising features. Figure 5.23 (right) highlights that already two well-chosen features are sufficient to effectively split the data into distinct clusters and that with an ensemble of linear classifiers these four activities could be separated well across the 14 study participants. The one outlier point from the post exercise activity (in orange) that appears within the relaxing (green) cluster was due to the algorithm failing to correctly estimate the respiration signal.

5.6.2 *E-Health and Telemedicine*

This section is based on the peer reviewed publication [21]. The idea to use the system in the context of e-health and telemedicine originates from me and I extracted the core elements of this idea from the publication. The study experiments were conducted by Steffen Brinkmann and the full evaluations and details can be found in the original work.

In this section, a respiration monitoring system for e-health and telemedicine applications is envisioned, with examples ranging from a personal usage for instance in sports or meditation, up to professional health monitoring, or beyond. The idea is that non-professionals should be able to operate such a system in an everyday environment. Thus, it should deliver reliable respiration data and at the same time be easy to set up and use. Since depth cameras already are deployed in many consumer grade devices, for instance in many modern smartphones, the usage of remote respiration estimation techniques could be viable to achieve this. A user in such a case simply has to sit or stand in front of the depth camera. When using the method proposed in Section 5.2, not even a clear line of sight to the user's upper body is required and motion artifacts caused by standing do not cause a significant challenge.

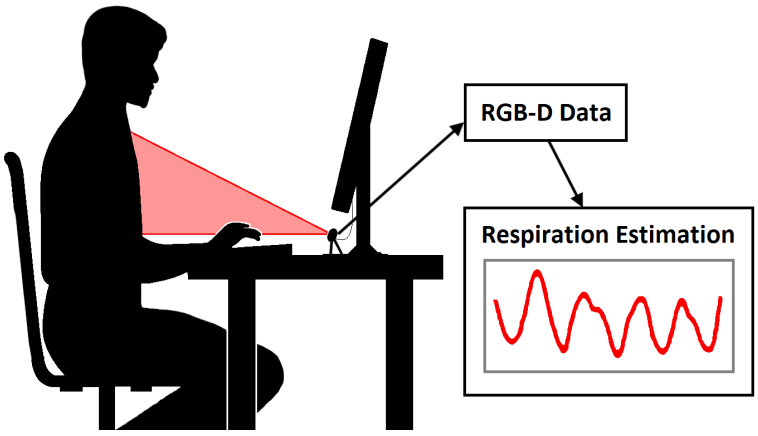

Figure 5.24: A user's chest is observed by an unobtrusive depth camera that is used to estimate and monitor this user's respiration.

To investigate such a scenario, a simple, yet expressive case study is conducted to demonstrate the feasibility of depth-based respiration estimation for e-health applications. In this study, an office-like setting is chosen, where a user sits in front of a display while being monitored by a depth camera that is installed like a small webcam, i.e. it has a clear line of sight to the user's upper body as depicted in Figure 5.24. This setup already resembles many use cases such a respiration sensor would be used in, be it to do some sort of breathing exercise or to just be monitored for a certain period of time. Possible applications could be a remote respiration measurement by a doctor in medical applications or a long-term respiration monitoring during screen work or while watching TV. This way, it can also be used to recognize problematic respiratory events at an early stage, for instance for asthma patients. The user furthermore could practice guided respiration exercises in front of a display that directly provides some sort of feedback or the user could share the respiration sig-

nal with a supervisor to get a more sophisticated feedback. Moreover, unobtrusively monitoring the respiration can serve as a fatigue indicator in safety critical functions, for instance in cockpits or control rooms.

The envisioned system was evaluated in a small-scale study on a group of 8 participants, which consists of 4 male and 4 female participants, all aged 21 to 57 years old.Each participant was recorded by a small consumer grade depth camera, namely the Intel RealSense D435 [66], that is attached to a standard consumer grade PC like a webcam. The D435 is smaller than the Kinect and has less demands on power consumption: It can directly be powered via the USB port. In contrast to the Kinect, furthermore, the computation of the depth values is not based on time of flight but on stereo imaging. Similar to the systematic parameter evaluation (see Section 5.3.3), the system was evaluated at different distances of 1, 2, and 3 meters, each at two respiratory rates of 10 bpm (0.25 Hz) and 15 bpm (0.17 Hz), respectively. The participants hereby are asked to sit upright in an office chair and follow a paced breathing visualization. The participants this time are not standing, nor are they allowed to occlude their upper body. To assess a ground truth signal, all participants furthermore are wearing a Vernier GoDirect respiration belt below their shirts or pullovers. Figure 5.25 depicts an exemplary depth frame taken with the D435 depth camera from a distance of 1 meter, with the region of interest to sample the respiration signal from overlaid as a red rectangle. An example of the respiration signals of both the depth camera and the respiration belt can be found in Figure 5.26, where both signals are overlaid for comparison reasons.

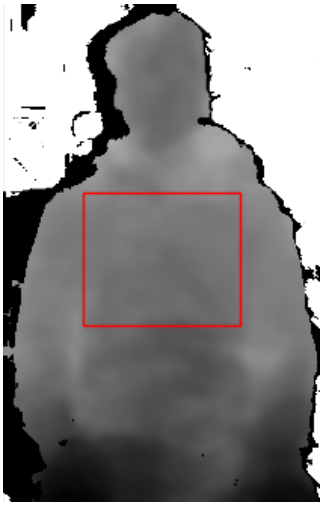

Figure 5.25: Exemplary depth frame at 1 m, color-coded in grey. The red rectangle indicates the region to sample the respiration signal from [21].

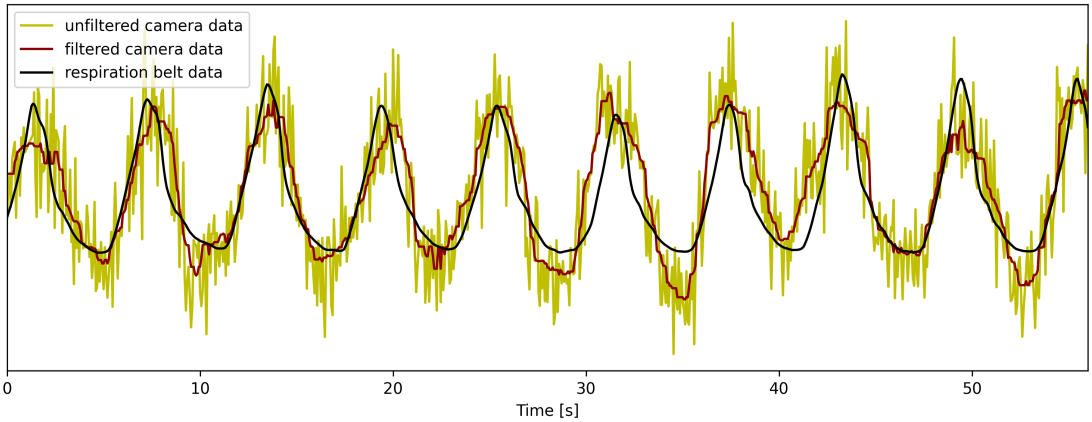

Figure 5.26: Example of the unfiltered (yellow), filtered (red), and ground truth (black) respiration signals of user 1 at 2 meters distance with a breathing rate of 10 bpm, as obtained from the depth camera or the respiration belt, respectively [21].

Results show that the proposed system is feasible to be used for daily life respiration estimation in the context of e-health. Especially at close distances and regardless of the gender, the respiration estimation worked reliably. Furthermore, participants were not required to adjust their clothes and performance values did not diverge for participants with loose-fitting textiles. The full user study can be found in [21].

The experiments performed in this work can only be considered a small-scale feasibility study. For a deployment in a medical context, a more thorough user study is indispensable. It should not only comprise more participants, but should also consider potential respiratory illnesses as well as a wider range of examined respiratory rates, including various other parameters.

5.7 DISCUSSION

In light of the above results, this section will discuss the limitations, assumptions and requirements for the data, the methods, and the evaluations.

5.7.1 *Limitations of the Dataset and Evaluation Setup*

The dataset was recorded with the intention to be as realistic as possible, yet it should be applicable to existing respiration estimation methods and it should allow a comparison of those among each other, independent of user-specific or external influences. Thus it was recorded under certain assumptions and with certain study design decisions that limit the applicability of the found results to more general scenarios. The assumptions and decisions being made are:

- The user generally faces the depth camera and only is rotated by a small amount to either side. Only a single user is recorded at the same time.
- The user is sitting or standing upright at a fixed position with a distance of 1, 2, 3, or 4 meters to the depth camera during the systematic parameter study and standing at a distance of 3 meters during the validation study. The user does not consciously lean excessively to either side, bend the upper body forwards or backwards, or moves towards the camera or to any other location. During the validation study, however, some participants did unconsciously bend the upper body forwards.
- Upper body motion is restricted to a small amount, like swaying while keeping balance, repositioning movements to either side, for instance, when switching from one leg to another, or small body rotations. Rotating the body actively away from the depth camera is not allowed.
- The user may occlude its upper body with one or both hands and with an in-hand object (a mug) arbitrarily during the occlusion condition.
- For the systematic parameter evaluation, users are adhering their respiratory rate to a breathing visualisation with fixed frequency. This is not a realistic setup, but eliminates the influence of user-specific breathing styles and paces.

Also a fixed respiratory rate makes the benchmarking of the different methods easier and better comparable, even across different users.

- The users are wearing a big variety of regular indoor clothing. To reflect more realistic indoor scenarios, users were not asked to wear specific cloths nor were the recordings repeated on a set of different clothing types. Some users, however, are wearing different clothing on different recordings. Also, yet there is no systematic classification of the clothing styles.

The datset, with 7 female and 17 male participants, is not balanced, so a comparison of both groups likely contains bias. User 9 had difficulties in adhering to the paced breathing setup and other users might occasionally also show deviations. Ground truth was only recorded explicitly for the validation study with a respiration belt and otherwise was obtained from the paced breathing. For the systematic parameter evaluation, accuracy, error, and signal-to-noise evaluations are obtained by comparing the measured respiratory rate to the ground truth frequency as given by the setup of the paced breathing visualization. The Pearson Correlation Coefficient is obtained by comparing the measured breathing signal to the ground truth signal from the respiration belt (validation study) or to a sine wave of the respective frequency (systematic parameter evaluation).

The FFT window has a fixed length of 40s for the validation study and 48s for the systematic parameter evaluation, and the window moves with a step size of 10s for the validation study or one breathing cycle (4s or 6s) as given by the respective breathing rate setup during the systematic parameter evaluation. For the validation study, furthermore the signals are band-pass filtered and a Hann-window is applied for the FFT. The performance values are likely to change with different FFT parameters.

5.7.2 *Limitations of the Proposed Method*

The proposed method relies heavily on tracking the torso and the differences between a region of the torso that is less affected by respiration at the throat, and a heavily respiration-affected area within the torso. As the throat's region is relatively small, it is susceptible to noise, occlusions, and clothing effects such as a moving collar. This makes the proposed method more susceptible in winter or colder climates, for instance, when users might wear heavier attire that completely covers the throat region, too. Another limitation is that the proposed model cannot deal with large rotational offsets and, at the moment, requires the user to keep a steady distance to the depth camera. The latter can be solved by rescaling the input frames or the model, large rotations on the other hand are more challenging. In a similar implementation, they would for instance require a three dimensional model and an iterative closest point algorithm to match the input frames to the body surface, making it hard to keep real-time performance especially on embedded platforms. Small rotational or spatial movements, however, are present throughout the whole dataset, especially during the standing postures. The proposed model can adapt to these and, due to the difference-based signal extraction, the proposed method performs remarkably well under these conditions. On a more general application, where users can be walking,

running (on a treadmill), or performing fast body movements, the proposed method yet has to be validated and will likely perform poorly.

Another issue is that in some cases low frequency components dominate in the frequency domain, even when a clear respiration signal is present in the time series. An alternative would be to use peak detection or a zero crossings method to find the respiratory rate more timely and without the effect of lower frequencies showing up. These methods on the other hand have to deal with noise that especially at higher distances significantly increases and may have decreased precision due to false positives or signal distortions. Different types of environmental noise can thus impact the quality of the proposed method's resulting respiration signal. It is possible that adaptive digital filtering noise cancellation techniques be integrated to provide more accurate results in future versions of the proposed method.

Since the depth imaging frames can contain a user anywhere in the frame, the proposed method is expected to work well with multiple users, as long as these do not block each other. This likely becomes challenging for detecting users' positions and joints in the depth frames accurately as more users are present. This multi-user scenario was not pursued further, as the focus of the evaluation was on assessing the influence of different parameters on depth-based respiration estimation methods on a per user basis and without complicating the evaluation unnecessarily. The proposed method, however, can easily be scaled up to multiple users, but nonetheless should be validated in a separate experiment then.

Finally, another limitation of the proposed method has to be mentioned: The surface reconstruction does in some cases not deliver good results, for instance if the model is initialized with occlusions, as illustrated in Figure 5.27 (left). It however does ensure that over time the recovered parts are elevated accordingly to the surrounding pixels while keeping sufficient surface detail to distinguish between occluded and non-occluded regions in the following input frames. Moreover the model is adaptive to small surface deformations and does not require a valid initialization. Any occlusions erroneously incorporated into the proposed model will fade away as soon as the occluding object moves away, leaving behind a better surface approximation as depicted in Figure 5.27 (right), but this does take several frames.

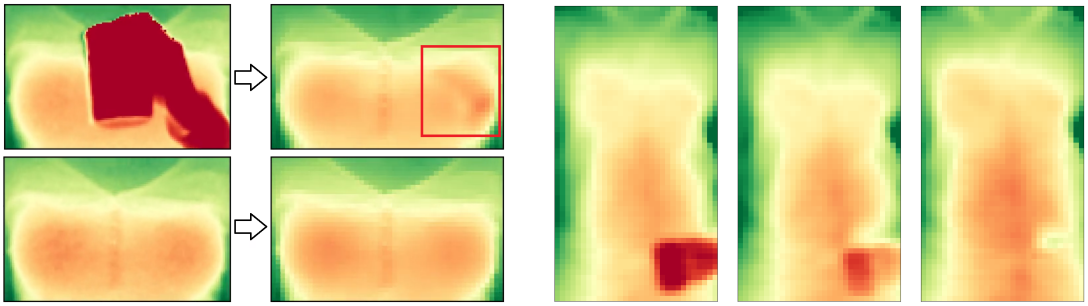

Figure 5.27: **Left**: Comparison of occlusion vs no occlusion on the prediction. Note the difference in the highlighted area. **Right**: Example of the adaptiveness of the model: An initial occlusion fades away as soon as the occluded area gets visible.

5.7.3 *Limitations of the Systematic Parameter Evaluation*

For the systematic parameter evaluation, the majority of current state-of-the-art methods were evaluated and compared under a variety of different settings (see Section 5.5). Current state-of-the-art methods are the *PCA*, *Mean Raw*, and *Diff Model* (the proposed method). Additionally, using the median instead of the mean was proposed, and the model based approach was leveraged to a more lightweight version where only the difference-based approach is used instead of computing a whole torso model. These modifications yield three more methods, namely the *Median Raw*, the *Diff Mean*, and the *Diff Median*.

One approach left out in the evaluations are volume-based methods. These first create a mesh model of the torso surface via triangulation and use it to compute the change of volume. Since the back of the torso is not visible to the depth camera, in related works, a certain, constant depth threshold is used to form a plane that bounds the mesh to the back. Any torso movement, be it respiration or usual body motion, thus will change the mesh volume. Bounding the mesh to a constant threshold at the back consequently is equivalent to computing a weighted sum of the depth values, similar to the mean-based methods. Using the edges of the torso or a defined torso area to compute the bounding plane at the back of the torso dynamically, on the other hand, is not a reliable process as it is subject to unreliable bounding window alignments, movements, surface deformations, and additionally is strongly correlated with breathing movements as was found in the experiments (also see Figure 5.6). With these restrictions, and from the findings that the volume-based approach is less accurate while being computationally much more expensive as conducted by [152], explicitly computing a mesh is omitted. Instead, the studies rely on computing the mean or the median of the torso depth pixels to approximate the change in torso elevation, which, when multiplied by the torso width and height, would yield a torso volume approximation, too. With the difference-based approaches, furthermore a dynamic threshold to the back is modelled that is able to leverage most of the motion artifacts entering the mean or median approximations. Although the state-of-the-art volume based approaches' performance can be considered to be in the range of the *Mean Raw*, explicitly modelling a 3D torso surface and fitting it to the depth data may have great potential for depth-based respiration estimation. Such a 3D model can potentially aid in overcoming many current limitations of depth-based remote respiration estimation methods. A systematic evaluation of the volume-based methods thus would be beneficial for extending these methods and for further research in this direction.

Some limitations apply to the *PCA* method. As suggested from the related work, and to achieve a run-time respiration estimation, the first 180 frames are used to build the *PCA* model. Since it is unclear which principal component to select algorithmically, only the one with the highest eigenvalue is used, making the *PCA* method susceptible to motion artifacts that happen within the first 180 frames. A solution could be to perform an offline *PCA* on the whole signal and to manually select the most reasonable component, but motion artifacts are likely to enter the estimated breathing signal anyway. Furthermore, [167] suggest to apply a varimax rotation to the *PCA* components to feature local deformations that differentiate be-

tween thoracic or abdominal breathing. Their study, as well as all other studies on *PCA* based methods, however, was performed on participants lying still in supine position and wearing tight clothing with no folds, thus letting the method only deal with the two breathing styles and noise. A varimax rotation is not applied in the *PCA* implementation for the evaluation since motion artifacts, clothing-related surface deformations, and occlusions are heavily present, which can be expected to pose the major limitation to all *PCA* based methods in more realistic scenarios, at least when it is not possible to carefully inspect and select the correct principal components manually. It also has to be mentioned that this method is computationally expensive and furthermore requires a certain amount of reliable training data at the beginning and for each user individually.

5.7.4 *Comparison of Depth-Based Respiration Estimation to Non-Depth-Based Approaches (Wearable or From a Distance)*

Wi-Fi-based respiratory rate detection is a promising method as it comes at almost no cost, given a Wi-Fi enabled infrastructure. According to the related work, the Wi-Fi antennas however usually require a particular setup and alignment towards the user [165]. The antennas in most cases are in front of a lying, sitting, or standing still person within the sensing range of about 1 to 3 m and, according to [176], respiration sensing tends to fail when the observed person performs hand gestures. It can be assumed that even small body movements while having to stay still for a while, as observed in the experiments, will cause noticeable signal distortions. If Wi-Fi-based respiration monitoring works in a realistic environment, i.e., sender and receiver not in the same room or not within a range of a few meters and the observed person being randomly aligned, and how well this method performs if not remains an unsolved research question. Furthermore, if two or more users are within the detection range, it remains unclear how to assign a respiration signal to the correct person or how to distinguish different breathing signals that are likely to interfere on the common carrier, especially when they overlap in frequency domain. A depth-based method such as the proposed has the benefit that users can be further away, can be distinguished and assigned the correct respiration signal that only is present on depth pixels covering a user's torso, and that changing environmental conditions, such as other persons walking through in the background of the scene, have less of an effect. Also, depth images provide a valuable insight in user movements that can be exploited to reduce motion artifacts. The fact that the depth camera can easily be located and blocked is both an advantage (transparency to users) and a disadvantage of the proposed method.

While special devices, ranging from spirometers to respiration belts, in general yield optimal results, they tend to be expensive and uncomfortable to wear for longer stretches of time. Furthermore, they have to be made available to the user and may need a supervisor for the setup. Having to wear a mask or respiration belt that might need occasional readjusting (due to it being too tight or too loose) from time to time certainly imposes distraction to the user. Other, less distracting wearable devices such as PPG equipped smart watches or fitness bands also have to be available to the user and need to be connected to the monitoring system. In applications where

users for a particular activity recognition application are not willing or not able to wear on-body sensors, a depth-based method has been found a viable alternative.

5.8 SUMMARY

Estimation of respiratory rate from depth data becomes significantly harder when the observed person is standing freely. When, as in prior work, persons are lying down in a supine position or sitting in a chair, body movement is restricted significantly. The studies performed here showed that even when persons are standing in front of a depth camera, traditional optical respiration estimation approaches tend to fail when (1) the body sways through motion from hips and legs, as well as (2) occlusions, in particular those coming from the persons gesturing themselves. Previous approaches show here significant drops (to around 20%) in accuracy.

For this reason, a novel approach to remotely monitor the respiration from users facing a depth camera was presented. The proposed approach focuses on robustly segmenting the data from the user's torso in the depth images using the detected user's body joints and modeling this torso area over time. The respiration signal is obtained by a difference based approach, where a motion reference signal from a barely respiration affected body region is subtracted from the signal obtained from a body region that is heavily affected by respiration, i.e. from within the torso surface model. It was shown that it is possible to detect the breathing rate, even when the observed person is standing upright and occasionally occludes its torso. The proposed method furthermore is fairly light-weight and is able to run in real-time, possibly even on embedded platforms.

Two sets of experiments were performed collecting and analyzing data from 24 study participants, validating the proposed approach with a commercial wearable respiration monitor and examining crucial parameters for depth-based respiration estimation methods in general, such as different conditions (sitting, standing, and standing while performing drinking gestures) or different distances from 1 to 4 meters. The validation study confirmed that the proposed method does work as intended and for most users and activities, it has an accuracy that is comparable to a respiration belt. Key findings of the systematic parameter evaluation can be summarized as:

- The observed torso region influences both performance and signal quality for all methods: Under all circumstances, the results confirm that the chest is the ideal region for capturing the respiration signal. The abdomen region yields the lowest performance and signal quality, especially in the standing and occlusion scenarios.

- User condition (sitting, standing, or occluding their torso) affect performance and signal quality significantly for all methods. Non-difference-based methods tend to fail when persons are standing or when they move their arms in front of their torso. When users are standing, all difference-based methods show good performance values. In the presence of occlusions, the *Diff Model* and the methods that use the median to compute the respiration signal are recommendable.

- Different users deliver varying qualities of breathing signals, with few users performing significantly worse than most other users. Some users move a lot, longer hair can be a problem, and clothing can play a role: Some clothing poorly reflects infrared light, some garments have ribbons in front of the chest that interfere with the breathing detection. Difference-based methods have difficulties when the throat area is covered by a collar that moves while breathing. For some users, the respiration signal estimation works better than on all other users, regardless of the method used, for instance due to the ability of these users to stand still for a while or due to a different breathing style.

Other parameters were found to play a minor role. The distance between user and depth camera has less influence on performance, but a strong influence on the signal quality. Optimal distances are in the range of 1 to 2 meters, with higher distances causing more noise in the respiration signals. During occlusions, 2 meters led to the better results. The respiratory rate has only little effect: Higher rates are easier to detect, likely due to more breathing periods falling within a fixed-length FFT window. The signal quality for the higher respiratory rates was over all methods slightly reduced, though. Gender-dependent differences in the respiration estimation are due to unbalanced and the rather small group sizes hard to interpret.

The proposed method (*Diff Model*) showed best accuracy and signal quality results across all scenarios. In some use cases, however, other methods do have their benefits: If users are sitting, the non-difference-based methods perform equally well and only show a slightly decreased signal quality. The *Mean Raw* and the *Median Raw* hereby benefit from being computationally much less expensive and do not require a fixed size of the torso window. When the user moves closer or further away from the depth camera, these methods do not need to reinitialize a model. The same applies to the *Diff Mean* and *Diff Median* when users are standing. Using the median hereby has been shown to be superior to using the mean for extracting the breathing signal. *PCA* does not yield better performance values than the *Median Raw* and is about in the range of the *Mean Raw*, but requires an expensive training phase that is susceptible to any deformation or movement larger than or in the range of the breathing related chest or torso expansion. Using *PCA* thus should only be considered for use cases with tight clothing and no body movements, and where for instance a detailed torso surface model needs to be reconstructed. In use cases with negligible body motion and no occlusion, like in a sitting condition, and especially when computation time is limited like on an embedded system, the use of the *Median Raw* is recommended. The same applies to the *Diff Median* in the case of a scenario with motion artifacts, for instance when persons are standing. The breathing signal in this case should be low-pass filtered, especially on higher distances. Using the *Diff Model* during standing as well as in the presence of occlusions, however, yields better results.

In two case studies, it was shown that the respiration signal as obtained from depth-based respiration estimation can be used in applications such as e-health and telemedicine and that it is suitable to be used as an additional modality for activity recognition purposes, where the signal was used to distinguish between activities with otherwise high similarity. Recalling the complementary sensing approach from Chapter 3, the remotely captured respiration signal can well be sent to a body-worn

device where it complements other measured physiological signals for instance to be used for fitness tracking or to aid in activity recognition.

The anonymized dataset with depth data and respective body joints locations, as well as the proposed method's source code and the python experiment scripts that were used for validating the proposed method are available to support the reproduction of the method and results, and can be obtained by contacting the author or visiting https://ubicomp.eti.uni-siegen.de/home/datasets/sar20 and https://ubicomp.eti.uni-siegen.de/home/datasets/fcs21.

All subjects gave their informed consent for inclusion before they participated in the studies. The studies were conducted in accordance with the Declaration of Helsinki, and were approved by the Ethics Committee of the University of Siegen (ethics vote #ER_12_2019).

6 CONCLUSION

This dissertation presented an in-depth analysis of the different types of human body motion present in depth data, ranging from large-scale limb movements down to subtle movements of the upper body during respiration. It was shown how this data can be used for different applications and use-cases, for instance to achieve complementary motion sensing or to monitor a person's respiration from a distance.

More specifically, in various experiments and user studies, it was shown how optical and inertial motion data can be combined to achieve complementary motion sensing, how the person and the limb an inertial sensing device is worn on can be identified within this complementary data, and how such human motion data can efficiently be compressed using Piecewise Linear Approximation (PLA). Furthermore, it was shown how respiration can remotely be obtained from depth data in a robust way and what parameters influence such a respiration measurement.

While these research parts were addressed on their own, they all integrate well into a coherent picture. All these research fields, although seemingly being disparate, are linked by the principle of working with human body motion and can seamlessly be integrated into a wider complementary motion sensing approach where data acquired from a personal wearable device can be augmented with externally measured respiration data. Such an approach can have benefits in many scenarios. When used in sports or meditation applications, but also in medical applications and e-health scenarios, a more holistic image of the state of an observed user can be drawn. At the same time, the user has immediate access to this data via its wearable device. Also, from a more general perspective, such a data augmentation can lead to improvements in activity recognition, or it can simply be used to further optimize person identification by comparing respiration signals obtained from different modalities.

As outlined in this dissertation, such a wider complementary motion sensing approach can be achieved by pursuing the following steps: First, human body motion, including respiration, is measured from all observed users by an external depth camera. Simultaneously, the wearable devices of all users measure the movements of the limb they are attached to and transmit this data to the external system, given this service is enabled by the respective user. In order to save bandwidth and reduce energy consumption on mobile devices, this motion data can be compressed using PLA. On the external system, the motion data acquired from all wearable devices then can be matched to the motion data of the depth camera. As a result, each person's wearable device can be affiliated with externally measured data specific to this person, i.e. this person's respiration, but also other person related data such as posture or position. This data then is sent back to the respective device where it can be used in a range of existing applications or even lead to the emergence of applications that would not be possible without such data. This, however, is something that needs to be investigated in the future.

The core contributions and most important results of this dissertation that further the current state of the art can be listed as follows:

- The necessary steps and considerations that are required to achieve complementary motion sensing from optical and inertial Motion Capturing (MoCap) data were investigated. The benefits of such a complementary approach were demonstrated in a case study with 10 participants, where a combined use of both modalities improved the tracking performance of the participants' wrists.

- A novel method was proposed that enables the identification of the person and the limb an inertial sensing device is worn on within such complementary motion data. It is based on comparing and matching limb movements from both modalities, with results of the evaluation showing that the correct person and limb can be identified within 2.5 to 4.3 seconds. The core contribution hereby lies on the idea of the method itself. It can be considered an important algorithmic component for combining inertial and optical motion data that enables a variety of applications beyond mere motion capturing, for instance indoor localization on a wearable device or person reidentification in video streams. So far, only few coarsely related approaches can be found in the literature and none of them are intended to identify a person or limb in the first place. The proposed method thus closes a gap in this research.

- Piecewise Linear Approximation (PLA) has been found to be a suitable compression scheme for quaternion-based motion data. The idea behind using PLA is that similar to using keyframes in computer animation, only data points that mark a change of a motion need to be stored while data in between can be approximated through interpolation. Consequently, the key requirements for applying PLA algorithms to quaternion-based motion data were analyzed, with the outcome that such an algorithm needs to produce connected segments where the produced segment points have to be a subset of the original data. One reason is that otherwise all quaternions of the compressed signal would need to be normalized again. Another reason, however, is of much higher importance: Even slight deviations of the segment points from the original data can lead to significant angular deviations. This especially is true if different deviations occur on different axes. So far, no efficient PLA method, i.e. a PLA method with a constant time and memory complexity with respect to the compression ratio, exists that meets these constraints. Based on the above analysis, the novel method fastSW was proposed, which closes this gap in the state of the art. fastSW is specifically tailored for the processing of quaternion data and can directly be deployed on a sensor node or similar environments with limited computational resources.

- Depth-based respiration estimation methods in prior work require persons to lie down in a supine position or to sit still in a chair in order to restrict body movement to a minimum. When, in contrast to that, the person to be observed is standing freely, prior methods tend to fail because an estimation of a person's respiration becomes significantly harder. This situation becomes even worse in the presence of occlusions, for instance when the

observed person performs hand gestures in front of its torso. Experiments in this work revealed that the accuracy in such a case can drop down to about 20% for some methods. For this reason, a novel, depth-based method for the remote and contact-less monitoring of human respiration was proposed. The novel method allows a user to stand freely and is robust against partial occlusions of the user's upper body. Thus, the proposed method can be considered the most stable depth-based respiration estimation method to be found in the current state of the art. This stability opens up a range of new use-cases for respiration estimation, especially in home applications or applications in daily living, with examples ranging from meditation and fitness monitoring over activity recognition up to e-health and telemedicine. The proposed method thus can be considered an important contribution to the state of the art. The proposed method's source code can be found online by visiting https://ubicomp.eti.uni-siegen.de/home/datasets/sar20 (the method's script files are located within the validation dataset).

- So far, there is not a single publicly available dataset on depth-based respiration estimation. Moreover, prior studies were conducted under specific assumptions and contain little variability due to a modest number of participants and a limited, often specific set of parameters and conditions that were evaluated. In almost all studies, participants were required to lie down or to remain sedentary, often even had to wear tight clothing, and in the rare examples where participants were allowed to move, this movement still is heavily constrained. In fact, there is no study where participants were allowed to stand freely or even to occlude their upper body. Also, the influence of different parameters is neglected in prior work and the only study that is evaluated on different parameters (i.e. on sampling rate, user orientation, and clothing) is missing any kind of investigation of these parameters. To close this gap in the research, two publicly available datasets were recorded. One dataset is intended for the validation of depth-based respiration estimation methods by comparing it to a commercial wearable respiration belt, and the other one aims at a systematic evaluation of crucial parameters of these methods. In total, both datasets comprise 422 unique recordings from 24 different participants, where each participant was recorded from different distances, respiratory rates, and user activities, summing up to a combined length of more than 11 hours. The anonymized datsets as well as the python experiment scripts can be found online by visiting https://ubicomp.eti.uni-siegen.de/home/datasets/sar20 for the validation and https://ubicomp.eti.uni-siegen.de/home/datasets/fcs21 for the systematic parameter study.

- As already stated above, the influences of a variety of important parameters that affect the depth-based estimation of human respiration remain unknown. For this reason, an in-depth evaluation of to-date unknown influences of important key parameters on the most common state-of-the-art depth-based respiration estimation methods were conducted. Examined parameters are the observed torso region, whether the user is sitting, standing, or standing with regular self-occlusions, the distance to the depth camera, the respiratory rate,

the gender, and user-specific influences. The evaluation aims at revealing the strengths and weaknesses of the most common state-of-the-art methods with respect to the various parameters, but also allows a comparison of the different approaches among each other. Such a comparison, let alone such a deep insight into the single methods' performances under various parameters was not possible before and likely aids in future research and decision making.

- Finally, possible applications for depth-based respiration estimation are explored. In a first use-case, the applicability of using respiration as an additional modality in activity recognition was shown, and in a second use-case, its general suitability for e-health and telemedicine was demonstrated. Both are intended as feasibility studies with the aim to show up possible applications and the hope that following research takes up these approaches and that real applications will emerge in the future.

Although complementary motion sensing and remote respiration estimation are promising concepts, their use in applications is not widely seen. For complementary motion sensing, a likely reason might be that it is not straightforward to achieve or that example usages or simply a link from purely research oriented applications to real world use cases is missing. Future research should therefore focus more on the investigation of scenarios where complementary motion sensing is reasonable and on what could be achieved in such scenarios beyond mere motion capturing. In general, it has the potential to extend into many other fields and applications and have a significant influence there. Examples range from indoor navigation and data augmentation on mobile devices over AR, VR, and future user interfaces up to enhanced activity recognition or even applications in healthcare.

Remote respiration estimation on the other hand still is underrepresented in research and many parameters and conditions as well as their influence remain unknown. This makes it hard to be used in a reliable way in unknown situations, especially when used in a context like healthcare or similar. Open challenges that so far remain unsolved range from scenarios where users move their upper body a lot, perhaps even at relatively high velocities like when running on a treadmill, over situations where users bend their upper body or rotate it away from the camera, as for instance might happen during fitness or meditation exercises, up to scenarios where users are allowed to walk around or even occasionally leave the field of view of the observing depth camera. Furthermore, user-specific parameters such as different clothing or individual breathing styles as well as irregular breathing due to illness need to be investigated. Apart from these, a variety of other unforeseen challenges might be waiting for future research since remote respiration estimation can be considered to be rather at the beginning than on the end.

To conclude, complementary motion sensing and remote respiration estimation already on their own provide a broad variety of research opportunities that all have the potential to lead to promising applications in the future and using both in combination even extends this space.

BIBLIOGRAPHY

[1] URL: https://www.vicon.com/ (visited on 07/11/2023).

[2] URL: https://www.optitrack.com/ (visited on 07/11/2023).

[3] Heba Abdelnasser, Khaled A. Harras, and Moustafa Youssef. "UbiBreathe: A Ubiquitous non-Invasive WiFi-based Breathing Estimator." In: *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. MobiHoc '15. New York, NY, USA: ACM, 2015, pp. 277–286. ISBN: 978-1-4503-3489-1. DOI: 10.1145/2746285.2755969. URL: http://doi.acm.org/10.1145/2746285.2755969.

[4] Anthony P Addison, Paul S Addison, Philip Smit, Dominique Jacquel, and Ulf R Borg. "Noncontact Respiratory Monitoring Using Depth Sensing Cameras: A Review of Current Literature." In: *Sensors* 21.4 (2021), p. 1135.

[5] F. Q. Al-Khalidi, R. Saatchi, D. Burke, H. Elphick, and S. Tan. "Respiration rate monitoring methods: a review." In: *Pediatric pulmonology* 46.6 (2011), pp. 523–529. DOI: 10.1002/ppul.21416. (Visited on 04/21/2022).

[6] Ali Al-Naji, Kim Gibson, Sang-Heon Lee, and Javaan Chahl. "Real Time Apnoea Monitoring of Children Using the Microsoft Kinect Sensor: A Pilot Study." In: *Sensors (Basel, Switzerland)* 17.2 (2017). DOI: 10.3390/s17020286. (Visited on 07/14/2022).

[7] Navid Amini, Majid Sarrafzadeh, Alireza Vahdatpour, and Wenyao Xu. "Accelerometer-based on-body sensor localization for health and medical monitoring applications." In: *PUC* 7.6 (2011), pp. 746–760.

[8] Hirooki Aoki, Masaki Miyazaki, Hidetoshi Nakamura, Ryo Furukawa, Ryusuke Sagawa, and Hiroshi Kawasaki. "Non-contact respiration measurement using structured light 3-d sensor." In: *SICE Annual Conference (SICE), 2012 Proceedings of*. IEEE. 2012, pp. 614–618.

[9] Hirooki Aoki and Hidetoshi Nakamura. "Non-Contact Respiration Measurement during Exercise Tolerance Test by Using Kinect Sensor." In: *Sports* 6.1 (2018), p. 23.

[10] Hirooki Aoki, Hidetoshi Nakamura, Kengo Fumoto, Kunihisa Nakahara, and Masaru Teraoka. "Basic study on non-contact respiration measurement during exercise tolerance test by using kinect sensor." In: *System Integration (SII), 2015 IEEE/SICE International Symposium on*. IEEE. 2015, pp. 217–222.

[11] Arash Atrsaei, Hassan Salarieh, and Aria Alasty. "Human arm motion tracking by orientation-based fusion of inertial sensors and Kinect using unscented Kalman filter." In: *Journal of biomechanical engineering* 138.9 (2016), p. 091005.

[12] Andreas Baak, Thomas Helten, Meinard Müller, Gerard Pons-Moll, Bodo Rosenhahn, and Hans-Peter Seidel. "Analyzing and evaluating markerless motion tracking using inertial sensors." In: *European conference on computer vision*. Springer. 2010, pp. 139–152.

[13] Sebastian Bauer, Benjamin Berkels, Joachim Hornegger, and Martin Rumpf. "Joint ToF image denoising and registration with a CT surface in radiation therapy." In: *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer. 2011, pp. 98–109.

[14] Sebastian Bauer, Jakob Wasza, and Joachim Hornegger. "Photometric estimation of 3D surface motion fields for respiration management." In: *Bildverarbeitung für die Medizin 2012*. Springer, 2012, pp. 105–110.

[15] Flavia Benetazzo, Alessandro Freddi, Andrea Monteriù, and Sauro Longhi. "Respiratory rate detection algorithm based on RGB-D camera: theoretical background and experimental results." In: *Healthcare technology letters* 1.3 (2014), pp. 81–86.

[16] Eugen Berlin and Kristof Van Laerhoven. "An on-line piecewise linear approximation technique for wireless sensor networks." In: *IEEE Local Computer Network Conference*. IEEE. 2010, pp. 905–912.

[17] Eugen Berlin and Kristof Van Laerhoven. "Detecting leisure activities with dense motif discovery." In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. 2012, pp. 250–259.

[18] Alexandre Bernardino, Christian Vismara, Sergi Bermudez i Badia, Élvio Gouveia, Fátima Baptista, Filomena Carnide, Simão Oom, and Hugo Gamboa. "A dataset for the automatic assessment of functional senior fitness tests using kinect and physiological sensors." In: *2016 1st International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*. IEEE. 2016, pp. 1–6.

[19] Matteo Bonini and Omar S. Usmani. "Novel methods for device and adherence monitoring in asthma." In: *Current Opinion in Pulmonary Medicine* 24.1 (2018), pp. 63–69. ISSN: 1070-5287. DOI: 10.1097/MCP.0000000000000439. URL: https://journals.lww.com/co-pulmonarymedicine/Fulltext/2018/01000/Novel_methods_for_device_and_adherence_monitoring.11.aspx.

[20] Marco Bortolini, Maurizio Faccio, Mauro Gamberi, and Francesco Pilati. "Motion Analysis System (MAS) for production and ergonomics assessment in the manufacturing processes." In: *Computers & Industrial Engineering* 139 (2020), p. 105485.

[21] Steffen Brinkmann, Jochen Kempfle, Kristof Van Laerhoven, and Jonas Pöhler. "Evaluation of a Depth Camera as e-Health Sensor for Contactless Respiration Monitoring." In: *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2023, pp. 136–141. DOI: 10.1109/PerComWorkshops56833.2023.10150271.

[22] Kim-Charline Broscheid, Sebastian Stoutz, C Chien-Hsi, and Lutz Schega. "The potential of a home-based gait evaluation system with a new low-cost IMU: A pilot study." In: *Conference: HEALTH ACROSS LIFESPAN (HAL)-International Conference on Healthiness and Fitness across the Lifespan. Magdeburg: Otto von Guericke University Magdeburg*. 2018, pp. 12–15.

[23] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods." In: *International journal of computer vision* 61.3 (2005), pp. 211–231.

[24] A. P. L. Bó, M. Hayashibe, and P. Poignet. "Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect." In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Aug. 2011, pp. 3479–3483. DOI: 10.1109/IEMBS.2011.6090940.

[25] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields." In: *IEEE transactions on pattern analysis and machine intelligence* 43.1 (2019), pp. 172–186.

[26] Diego Castro, William Coral, Camilo Rodriguez, Jose Cabra, and Julian Colorado. "Wearable-based human activity recognition using an iot approach." In: *Journal of Sensor and Actuator Networks* 6.4 (2017), p. 28.

[27] Fabio Centonze, Martin Schätz, Aleš Procházka, Jiří Kuchyňka, Oldřich Vyšata, Pavel Cejnar, and Martin Vališ. "Feature extraction using MS Kinect and data fusion in analysis of sleep disorders." In: *Computational Intelligence for Multimedia Understanding (IWCIM), 2015 International Workshop on*. IEEE. 2015, pp. 1–5.

[28] Hua-I Chang, Vivek Desai, Oscar Santana, Matthew Dempsey, Anchi Su, John Goodlad, Faraz Aghazadeh, and Gregory Pottie. "Opportunistic Calibration of Sensor Orientation Using the Kinect and Inertial Measurement Unit Sensor Fusion." In: *Proceedings of the Conference on Wireless Health*. WH '15. New York, NY, USA: ACM, 2015, 2:1–2:8. ISBN: 978-1-4503-3851-6. DOI: 10.1145/2811780.2811927. URL: http://doi.acm.org/10.1145/2811780.2811927.

[29] Charilaos Chourpiliadis and Abhishek Bhardwaj. "Physiology, Respiratory Rate." In: *StatPearls [Internet]*. Ed. by Charilaos Chourpiliadis and Abhishek Bhardwaj. StatPearls Publishing, 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK537306/.

[30] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "Improving human action recognition using fusion of depth camera and inertial sensors." In: *IEEE Transactions on Human-Machine Systems* 45.1 (2014), pp. 51–61.

[31] Shanshan Chen, John Lach, Benny Lo, and Guang-Zhong Yang. "Toward pervasive gait analysis with wearable sensors: A systematic review." In: *IEEE journal of biomedical and health informatics* 20.6 (2016), pp. 1521–1537.

[32] Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. "A review of hand gesture and sign language recognition techniques." In: *International Journal of Machine Learning and Cybernetics* 10 (2019), pp. 131–153.

[33] Carly Cooper, Anne Gross, Chad Brinkman, Ryan Pope, Kelli Allen, Susan Hastings, Bard E Bogen, and Adam P Goode. "The impact of wearable motion sensing technology on physical activity in older adults." In: *Experimental gerontology* 112 (2018), pp. 9–19.

[34] Michelle A Cretikos, Rinaldo Bellomo, Ken Hillman, Jack Chen, Simon Finfer, and Arthas Flabouris. "Respiratory rate: the neglected vital sign." In: *Medical Journal of Australia* 188.11 (2008), pp. 657–659.

[35] Antonio I Cuesta-Vargas, Alejandro Galán-Mercant, and Jonathan M Williams. "The use of inertial sensors system for human motion analysis." In: *Physical Therapy Reviews* 15.6 (2010), pp. 462–473.

[36] Andrea Giovanni Cutti, Gabriele Paolini, Marco Troncossi, Angelo Cappello, and Angelo Davalli. "Soft tissue artefact assessment in humeral axial rotation." In: *Gait & posture* 21.3 (2005), pp. 341–349.

[37] Massimiliano De Zambotti, Nicola Cellini, Aimee Goldstone, Ian M Colrain, and Fiona C Baker. "Wearable sleep technology in clinical and research settings." In: *Medicine and science in sports and exercise* 51.7 (2019), p. 1538.

[38] F. Destelle, A. Ahmadi, N. E. O'Connor, K. Moran, A. Chatzitofis, D. Zarpalas, and P. Daras. "Low-cost accurate skeleton tracking based on fusion of kinect and wearable inertial sensors." In: *2014 22nd European Signal Processing Conference (EUSIPCO)*. Sept. 2014, pp. 371–375.

[39] Elisa Digo, Mattia Antonelli, Valerio Cornagliotto, Stefano Pastorelli, and Laura Gastaldi. "Collection and analysis of human upper limbs motion features for collaborative robotic applications." In: *Robotics* 9.2 (2020), p. 33.

[40] Chendan Dou and Hao Huan. "Full Respiration Rate Monitoring Exploiting Doppler Information with Commodity Wi-Fi Devices." In: *Sensors (Basel, Switzerland)* 21.10 (2021). DOI: 10.3390/s21103505. (Visited on 04/22/2022).

[41] Amy D. Droitcour, Todd B. Seto, Byung-Kwon Park, Shuhei Yamada, Alex Vergara, Charles El Hourani, Tommy Shing, Andrea Yuen, Victor M. Lubecke, and Olga Boric-Lubecke. "Non-contact respiratory rate measurement validation for hospitalized patients." In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* 2009 (2009), pp. 4812–4815. ISSN: 2375-7477. DOI: 10.1109/IEMBS.2009.5332635. (Visited on 04/21/2022).

[42] Katrin Kroemer Elbert, Henrike B Kroemer, and Anne D Kroemer Hoffman. *Ergonomics: how to design for ease and efficiency*. Academic Press, 2018.

[43] Hazem Elmeleegy, Ahmed Elmagarmid, Emmanuel Cecchet, Walid G Aref, and Willy Zwaenepoel. "Online piece-wise linear approximation of numerical streams with precision guarantees." In: *Proceedings of the International Conference on Very Large Data Bases*. ACM. 2009.

[44] Heather E. Elphick, Abdulkadir Hamidu Alkali, Ruth K. Kingshott, Derek Burke, and Reza Saatchi. "Exploratory Study to Evaluate Respiratory Rate Using a Thermal Imaging Camera." In: *Respiration; international review of thoracic diseases* 97.3 (2019), pp. 205–212. DOI: 10.1159/000490546.

[45] Xenofon Fafoutis, Letizia Marchegiani, Atis Elsts, James Pope, Robert Piechocki, and Ian Craddock. "Extending the battery lifetime of wearable sensors with embedded machine learning." In: *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*. IEEE. 2018, pp. 269–274.

[46] Bingfei Fan, Qingguo Li, Tian Tan, Peiqi Kang, and Peter B. Shull. "Effects of IMU Sensor-to-Segment Misalignment and Orientation Error on 3-D Knee Joint Angle Estimation." In: *IEEE Sensors Journal* 22.3 (2022), pp. 2543–2552. DOI: 10.1109/JSEN.2021.3137305.

[47] Wei Fang, Lianyu Zheng, Huanjun Deng, and Hongbo Zhang. "Real-time motion tracking for mobile augmented/virtual reality using adaptive visual-inertial fusion." In: *Sensors* 17.5 (2017), p. 1037.

[48] Daniel Tik-Pui Fong and Yue-Yan Chan. "The Use of Wearable Inertial Motion Sensors in Human Lower Limb Biomechanics Studies: A Systematic Review." In: *Sensors* 10.12 (2010), pp. 11556–11565. ISSN: 1424-8220. DOI: 10.3390/s101211556. URL: https://www.mdpi.com/1424-8220/10/12/11556.

[49] Erich Fuchs, Thiemo Gruber, Jiri Nitschke, and Bernhard Sick. "Online segmentation of time series based on polynomial least-squares approximations." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.12 (2010), pp. 2232–2245.

[50] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. "Fusing visual and inertial sensors with semantics for 3d human pose estimation." In: *International Journal of Computer Vision* 127 (2019), pp. 381–397.

[51] Florian Grützmacher. "System-level design of energy-efficient sensor-based human activity recognition systems: a model-based approach." PhD thesis. Dissertation, Rostock, Universität Rostock, 2021. DOI: 10.18453/rosdok_id00003373.

[52] Florian Grützmacher, Benjamin Beichler, Albert Hein, Thomas Kirste, and Christian Haubelt. "Time and memory efficient online piecewise linear approximation of sensor signals." In: *Sensors* 18.6 (2018), p. 1672.

[53] Florian Grützmacher, Albert Hein, Benjamin Beichler, Polichronis Lepidis, Rainer Dorsch, Thomas Kirste, and Christian Haubelt. "Energy Efficient On-Sensor Processing for Online Activity Recognition." In: *Proceedings of the 8th International Joint Conference on Pervasive and Embedded Computing and Communication Systems*. 2018, pp. 223–230.

[54] Florian Grützmacher, Jochen Kempfle, Kristof Van Laerhoven, and Christian Haubelt. "fastsw: Efficient piecewise linear approximation of quaternion-based orientation sensor signals for motion capturing with wearable imus." In: *Sensors* 21.15 (2021). ISSN: 1424-8220. DOI: 10.3390/s21155180.

[55] Florian Grützmacher, Johann-Peter Wolff, Albert Hein, Polichronis Lepidis, Rainer Dorsch, Thomas Kirste, and Christian Haubelt. "Towards energy efficient sensor nodes for online activity recognition." In: *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*. IEEE. 2017, pp. 8291–8296.

[56] Federico Guede-Fernández, Mireya Fernández-Chimeno, Juan Ramos-Castro, and Miguel A García-González. "Driver drowsiness detection based on respiratory signal analysis." In: *IEEE access* 7 (2019), pp. 81826–81838.

[57] Zied Guendil, Zied Lachiri, Choubeila Maaoui, and Alain Pruski. "Multiresolution framework for emotion sensing in physiological signals." In: *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE. 2016, pp. 793–797.

[58] Guillermo Gutierrez, Jeffrey Williams, Ghadah A. Alrehaili, Anna McLean, Ramin Pirouz, Richard Amdur, Vivek Jain, Jalil Ahari, Amandeep Bawa, and Shawn Kimbro. "Respiratory rate variability in sleeping adults without obstructive sleep apnea." In: *Physiological reports* 4.17 (2016). DOI: 10.14814/phy2.12949. (Visited on 04/21/2022).

[59] Marian Haescher, Denys JC Matthies, John Trimpop, and Bodo Urban. "A study on measuring heart-and respiration-rate via wrist-worn accelerometer-based seismocardiography (SCG) in comparison to commonly applied technologies." In: *Proceedings of the 2nd international Workshop on Sensor-based Activity Recognition and Interaction*. 2015, pp. 1–6.

[60] Rabab A Hameed, Mohannad K Sabir, Mohammed A Fadhel, Omran Al-Shamma, and Laith Alzubaidi. "Human emotion classification based on respiration signal." In: *Proceedings of the International Conference on Information and Communication Technology*. 2019, pp. 239–245.

[61] Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. "Real-time body tracking with one depth camera and inertial sensors." In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 1105–1112.

[62] Roberto Henschel, Timo Von Marcard, and Bodo Rosenhahn. "Accurate long-term multiple people tracking using video and body-worn IMUs." In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8476–8489.

[63] Berthold KP Horn and Brian G Schunck. "Determining optical flow." In: *Artificial intelligence* 17.1-3 (1981), pp. 185–203.

[64] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. "Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time." In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 37.6 (Nov. 2018), 185:1–185:15.

[65] T. Igasaki, K. Nagasawa, I. A. Akbar, and N. Kubo. "Sleepiness classification by thoracic respiration using support vector machine." In: *2016 9th Biomedical Engineering International Conference (BMEiCON)*. 2016, pp. 1–5.

[66] Intel® RealSense™ Depth and Tracking Cameras. *Depth Camera D435*. 17.06.2021. URL: https://www.intelrealsense.com/depth-camera-d435/ (visited on 04/26/2022).

[67] Marco Iosa, Pietro Picerno, Stefano Paolucci, and Giovanni Morone. "Wearable inertial sensors for human movement analysis." In: *Expert review of medical devices* 13.7 (2016), pp. 641–659. DOI: 10.1080/17434440.2016.1198694.

[68] Delaram Jarchi, James Pope, Tracey KM Lee, Larisa Tamjidi, Amirhosein Mirzaei, and Saeid Sanei. "A review on accelerometry-based gait analysis and emerging clinical applications." In: *IEEE reviews in biomedical engineering* 11 (2018), pp. 177–194.

[69] P. Jatesiktat and W. T. Ang. "Recovery of forearm occluded trajectory in Kinect using a wrist-mounted Inertial Measurement Unit." In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. July 2017, pp. 807–812. DOI: 10.1109/EMBC.2017.8036947.

[70] Prayook Jatesiktat, Dollaporn Anopas, and Wei Tech Ang. "Personalized markerless upper-body tracking with a depth camera and wrist-worn inertial measurement units." In: *2018 40th Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 1–6.

[71] Wenchao Jiang and Zhaozheng Yin. "Combining passive visual cameras and active imu sensors to track cooperative people." In: *2015 18th International Conference on Information Fusion (Fusion)*. IEEE. 2015, pp. 1338–1345.

[72] Wenchao Jiang and Zhaozheng Yin. "Combining passive visual cameras and active IMU sensors for persistent pedestrian tracking." In: *Journal of Visual Communication and Image Representation* 48 (2017), pp. 419–431.

[73] Zygimantas Jocys, Arash Pour Yazdan Panah Kermani, and Daniel Roggen. "Multimodal fusion of IMUs and EPS body worn sensors for scratch recognition." In: *Proceedings of Pervasive Health 2020* (2020).

[74] Tomoya Kaichi, Tsubasa Maruyama, Mitsunori Tada, and Hideo Saito. "Resolving position ambiguity of imu-based human pose with a single rgb camera." In: *Sensors* 20.19 (2020), p. 5453.

[75] Christoph Kalkbrenner, Steffen Hacker, Maria-Elena Algorri, and Ronald Blechschmidt-Trapp. "Motion capturing with inertial measurement units and kinect." In: *Proc. Int. Joint Conf. Biomed. Eng. Syst. Technol*. Vol. 1. 2014, pp. 120–126.

[76] Paul J Keall, Gig S Mageras, James M Balter, Richard S Emery, Kenneth M Forster, Steve B Jiang, Jeffrey M Kapatoes, Daniel A Low, Martin J Murphy, Brad R Murray, et al. "The management of respiratory motion in radiation oncology report of AAPM Task Group 76." In: *Medical physics* 33.10 (2006), pp. 3874–3900.

[77] Jochen Kempfle and Kristof Van Laerhoven. "PresentPostures: A Wrist and Body Capture Approach for Augmenting Presentations." In: *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, Mar. 2018. DOI: 10.1109/percomw.2018.8480155.

[78] Jochen Kempfle and Kristof Van Laerhoven. "Human posture capture and editing from heterogeneous modalities." In: *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 489–494. ISBN: 9781450353786. DOI: 10.1145/3152832.3157812.

[79] Jochen Kempfle and Kristof Van Laerhoven. "Respiration Rate Estimation with Depth Cameras: An Evaluation of Parameters." In: *Proceedings of the 5th International Workshop on Sensor-Based Activity Recognition and Interaction*. iWOAR '18. New York, NY, USA: Association for Computing Machinery, 2018. ISBN: 9781450364874. DOI: 10.1145/3266157.3266208.

[80] Jochen Kempfle and Kristof Van Laerhoven. "Towards Breathing as a Sensing Modality in Depth-Based Activity Recognition." In: *Sensors* 20.14 (July 2020), p. 3884. ISSN: 1424-8220. DOI: 10.3390/s20143884.

[81] Jochen Kempfle and Kristof Van Laerhoven. "Breathing In-Depth: A Parametrization Study on RGB-D Respiration Extraction Methods." In: *Frontiers in Computer Science* 3 (2021). ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.757277.

[82] Jochen Kempfle and Kristof Van Laerhoven. "Quaterni-On: Calibration-free Matching of Wearable IMU Data to Joint Estimates of Ambient Cameras." In: *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE. 2021, pp. 611–616. DOI: 10/kt2k.

[83] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani. "An online algorithm for segmenting time series." In: *Proceedings 2001 IEEE international conference on data mining*. IEEE. 2001, pp. 289–296.

[84] H. Knutsson and C. . Westin. "Normalized and differential convolution." In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. June 1993, pp. 515–523. DOI: 10.1109/CVPR.1993.341081.

[85] Manon Kok, Jeroen D Hol, and Thomas B Schön. "An optimization-based approach to human body motion capture using inertial sensors." In: *IFAC Proceedings Volumes* 47.3 (2014), pp. 79–85.

[86] Eline van der Kruk and Marco M. Reijne. "Accuracy of human motion capture systems for sport applications; state-of-the-art review." In: *European Journal of Sport Science* 18.6 (2018), pp. 806–819. DOI: 10.1080/17461391.2018.1463397. eprint: https://doi.org/10.1080/17461391.2018.1463397. URL: https://doi.org/10.1080/17461391.2018.1463397.

[87] Kai Kunze and Paul Lukowicz. "Using acceleration signatures from everyday activities for on-body device location." In: (2007), pp. 115–116.

[88] Kai Kunze and Paul Lukowicz. "Sensor placement variations in wearable activity recognition." In: *IEEE Pervasive Computing* 13.4 (2014), pp. 32–41.

[89] Kai Kunze, Paul Lukowicz, Holger Junker, and Gerhard Tröster. "Where am i: Recognizing on-body positions of wearable sensors." In: (2005), pp. 264–275.

[90] Yung-Ming Kuo, Jiann-Shu Lee, and Pau-Choo Chung. "A visual context-awareness-based sleeping-respiration measurement system." In: *IEEE Transactions on Information Technology in Biomedicine* 14.2 (2010), pp. 255–265.

[91] HyeokHyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D. Abowd, Nicholas D. Lane, and Thomas Ploetz. "IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition." In: *CoRR* abs/2006.05675 (2020). arXiv: 2006.05675. URL: https://arxiv.org/abs/2006.05675.

[92] Oscar D Lara, Alfredo J Pérez, Miguel A Labrador, and José D Posada. "Centinela: A human activity recognition system based on acceleration and vital sign data." In: *Pervasive and mobile computing* 8.5 (2012), pp. 717–729.

[93] Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. "Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact." In: *Science Robotics* 6.58 (2021), eabe1315.

[94] Daniel Lemire. "A better alternative to piecewise linear time series segmentation." In: *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM. 2007, pp. 545–550.

[95] Tong Li and Haoyong Yu. "Upper Body Pose Estimation Using a Visual–Inertial Sensor System With Automatic Sensor-to-Segment Calibration." In: *IEEE Sensors Journal* 23.6 (2023), pp. 6292–6302.

[96] Tong Li and Haoyong Yu. "Visual-Inertial Fusion-Based Human Pose Estimation: A Review." In: *IEEE Transactions on Instrumentation and Measurement* 72 (2023), pp. 1–16. DOI: 10.1109/TIM.2023.3286000.

[97] Gabriele Ligorio and Angelo Sabatini. "Dealing with Magnetic Disturbances in Human Motion Capture: A Survey of Techniques." In: *Micromachines* 7.3 (2016), p. 43.

[98] WS Lim, SM Carty, JT Macfarlane, RE Anthony, J Christian, KS Dakin, and PM Dennis. "Respiratory rate measurement in adults—how reliable is it?" In: *Respiratory medicine* 96.1 (2002), pp. 31–33.

[99] Chong Liu, Kui Wu, and Jian Pei. "An energy-efficient data collection framework for wireless sensor networks by exploiting spatiotemporal correlation." In: *IEEE transactions on parallel and distributed systems* 18.7 (2007), pp. 1010–1023.

[100] Haipeng Liu, John Allen, Dingchang Zheng, and Fei Chen. "Recent development of respiratory rate measurement technologies." In: *Physiological measurement* 40.7 (2019), 07TR01. DOI: 10.1088/1361-6579/ab299e. (Visited on 07/15/2022).

[101] Bruce D Lucas, Takeo Kanade, et al. "An iterative image registration technique with an application to stereo vision." In: (1981).

[102] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. "Mediapipe: A framework for building perception pipelines." In: *arXiv preprint arXiv:1906.08172* (2019).

[103] Ge Luo, Ke Yi, Siu-Wing Cheng, Zhenguo Li, Wei Fan, Cheng He, and Yadong Mu. "Piecewise linear approximation of streaming time series data with max-error guarantees." In: *2015 IEEE 31st international conference on data engineering*. IEEE. 2015, pp. 173–184.

[104] Mitja Luštrek, Božidara Cvetković, Violeta Mirchevska, Özgür Kafalı, Alfonso E Romero, and Kostas Stathis. "Recognising lifestyle activities of diabetic patients with a smartphone." In: *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE. 2015, pp. 317–324.

[105] Sebastian Madgwick. "An efficient orientation filter for inertial and inertial/-magnetic sensor arrays." In: (2010).

[106] Sharmin Majumder and Nasser Kehtarnavaz. "Vision and inertial sensing fusion for human action recognition: A review." In: *IEEE Sensors Journal* 21.3 (2020), pp. 2454–2467.

[107] Yuichi Maki, Shingo Kagami, and Koichi Hashimoto. "Accelerometer detection in a camera view based on feature point tracking." In: *IEEE/SICE International Symposium on System Integration*. 2010, pp. 448–453.

[108] Randa Mallat, Vincent Bonnet, Mohamad Ali Khalil, and Samer Mohammed. "Upper limbs kinematics estimation using affordable visual-inertial sensors." In: *IEEE Transactions on Automation Science and Engineering* 19.1 (2020), pp. 207–217.

[109] Charles Malleson, John Collomosse, and Adrian Hilton. "Real-time multi-person motion capture from multi-view video and IMUs." In: *International Journal of Computer Vision* 128 (2020), pp. 1594–1611.

[110] Charles Malleson, Andrew Gilbert, Matthew Trumble, John Collomosse, Adrian Hilton, and Marco Volino. "Real-time full-body motion capture from video and imus." In: *2017 International Conference on 3D Vision (3DV)*. IEEE. 2017, pp. 449–457.

[111] Andrea Mannini, Angelo M Sabatini, and Stephen S Intille. "Accelerometry-based recognition of the placement sites of a wearable sensor." In: *Pervasive and mobile computing* 21 (2015), pp. 62–74.

[112] TV Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. "Multimodal motion capture dataset TNT15." In: *Leibniz Univ. Hannover, Hanover, Germany, and Max Planck for Intelligent Systems, Tübingen, Germany, Tech. Rep* (2016).

[113] Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. "Human Pose Estimation from Video and IMUs." In: *Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2016), pp. 1533–1547.

[114] Manuel Martinez and Rainer Stiefelhagen. "Breath rate monitoring during sleep using near-IR imagery and PCA." In: *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE. 2012, pp. 3472–3475.

[115] Carlo Massaroni, Andrea Nicolo, Massimo Sacchetti, and Emiliano Schena. "Contactless methods for measuring respiratory rate: A review." In: *IEEE Sensors Journal* (2020).

[116] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. "XNect: Real-time multi-person 3D motion capture with a single RGB camera." In: *ACM Transactions on Graphics (TOG)* 39.4 (2020), pp. 82–1.

[117] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera." In: vol. 36. 4. July 2017. DOI: 10.1145/3072959.3073596. URL: http://gvv.mpi-inf.mpg.de/projects/VNect/.

[118] Michael Christopher Melnychuk, Paul M. Dockree, Redmond G. O'Connell, Peter R. Murphy, Joshua H. Balsters, and Ian H. Robertson. "Coupling of respiration and attention via the locus coeruleus: Effects of meditation and pranayama." In: *Psychophysiology* 55.9 (2018), e13091. ISSN: 1469-8986. DOI: 10.1111/psyp.13091. URL: https://onlinelibrary.wiley.com/doi/10.1111/psyp.13091.

[119] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. "IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–12.

[120] Chaithanya Kumar Mummadi, Frederic Philips Peter Leo, Keshav Deep Verma, Shivaji Kasireddy, Philipp M Scholl, Jochen Kempfle, and Kristof Van Laerhoven. "Real-time and embedded detection of hand gestures with an IMU-based glove." In: *Informatics*. Vol. 5. 2. MDPI. 2018, p. 28. DOI: 10.3390/informatics5020028.

[121] Kazuki Nakajima, Yoshiaki Matsumoto, and Toshiyo Tamura. "Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed." In: *Physiological Measurement* 22.3 (2001), N21.

[122] Kazuki Nakajima, Atsushi Osa, and Hidetoshi Miike. "A method for measuring respiration and physical activity in bed by optical flow analysis." In: *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*. Vol. 5. IEEE. 1997, pp. 2054–2057.

[123] Richard A. Newcombe, Andrew J. Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. "KinectFusion: Real-time dense surface mapping and tracking." In: *2011 10th IEEE International Symposium on Mixed and Augmented Reality*. IEEE, Oct. 2011.

[124] Yuan Niu, Jinhua She, and Chi Xu. "A Survey on IMU-and-Vision-based Human Pose Estimation for Rehabilitation." In: *2022 41st Chinese Control Conference (CCC)*. IEEE. 2022, pp. 6410–6415.

[125] Philip J Noonan, Jon Howard, Deborah Tout, Ian Armstrong, Heather A Williams, Tim F Cootes, William A Hallett, and Rainer Hinz. "Accurate markerless respiratory tracking for gated whole body PET using the Microsoft Kinect." In: *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE*. IEEE. 2012, pp. 3973–3974.

[126] Štěpán Obdržálek, Gregorij Kurillo, Ferda Ofli, Ruzena Bajcsy, Edmund Seto, Holly Jimison, and Michael Pavel. "Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population." In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2012, pp. 1188–1193. DOI: 10.1109/EMBC.2012.6346149.

[127] D Ohlendorf, I Avaniadi, F Adjami, W Christian, C Doerry, V Fay, V Fisch, A Gerez, J Goecke, U Kaya, et al. "Standard values of the upper body posture in healthy adults with special regard to age, sex and BMI." In: *Scientific Reports* 13.1 (2023), p. 873.

[128] Karen Otte, Bastian Kayser, Sebastian Mansow-Model, Julius Verrel, Friedemann Paul, Alexander U. Brandt, and Tanja Schmitz-Hübsch. "Accuracy and Reliability of the Kinect Version 2 for Clinical Measurement of Motor Function." In: *PLOS ONE* 11.11 (Nov. 2016), pp. 1–17.

[129] Racheal Parkes. "Rate of respiration: the forgotten vital sign." In: *Emergency Nurse* 19.2 (May 2011), pp. 12–17. DOI: 10.7748/en2011.05.19.2.12.c8504. URL: https://doi.org/10.7748%2Fen2011.05.19.2.12.c8504.

[130] Jochen Penne, Christian Schaller, Joachim Hornegger, and Torsten Kuwert. "Robust real-time 3D respiratory motion detection using time-of-flight cameras." In: *International Journal of Computer Assisted Radiology and Surgery* 3.5 (Nov. 2008), pp. 427–431. ISSN: 1861-6429. DOI: 10.1007/s11548-008-0245-2. URL: https://doi.org/10.1007/s11548-008-0245-2.

[131] Alexandra Pfister, Alexandre M. West, Shaw Bronner, and Jack Adam Noah. "Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis." In: *Journal of Medical Engineering & Technology* 38.5 (2014), pp. 274–280. DOI: 10.3109/03091902.2014.909540. eprint: http://dx.doi.org/10.3109/03091902.2014.909540. URL: http://dx.doi.org/10.3109/03091902.2014.909540.

[132] Ngoc Duy Pham, Trong Duc Le, and Hyunseung Choo. "Enhance exploring temporal correlation for data collection in WSNs." In: *2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies*. IEEE. 2008, pp. 204–208.

[133] Pietro Picerno, Marco Iosa, Clive D'Souza, Maria Grazia Benedetti, Stefano Paolucci, and Giovanni Morone. "Wearable inertial sensors for human movement analysis: A five-year update." In: *Expert review of medical devices* 18.sup1 (2021), pp. 79–94. DOI: 10.1080/17434440.2021.1988849.

[134] Marco A. F. Pimentel, Alistair E. W. Johnson, Peter H. Charlton, Drew Birrenkott, Peter J. Watkinson, Lionel Tarassenko, and David A. Clifton. "Toward a Robust Estimation of Respiratory Rate From Pulse Oximeters." In: *IEEE transactions on bio-medical engineering* 64.8 (2017), pp. 1914–1923. DOI: 10.1109/TBME.2016.2613124. (Visited on 04/22/2022).

[135] Ronald Poppe. "Vision-based Human Motion Analysis: An Overview." In: *Comput. Vis. Image Underst.* 108.1-2 (2007), pp. 4–18. ISSN: 1077-3142.

[136] Sasanka Potluri, Arvind Beerjapalli Chandran, Christian Diedrich, and Lutz Schega. "Machine learning based human gait segmentation with wearable sensor platform." In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2019, pp. 588–594.

[137] Aleš Procházka, Martin Schätz, Oldřich Vyšata, and Martin Vališ. "Microsoft kinect visual and depth sensors for breathing and heart rate analysis." In: *Sensors* 16.7 (2016), p. 996.

[138] Ariana Tulus Purnomo, Ding-Bing Lin, Tjahjo Adiprabowo, and Willy Fitra Hendria. "Non-Contact Monitoring and Classification of Breathing Pattern for the Supervision of People Infected by COVID-19." In: *Sensors (Basel, Switzerland)* 21.9 (2021). DOI: 10.3390/s21093172. (Visited on 04/22/2022).

[139] Joern Rehder and Roland Siegwart. "Camera/IMU calibration revisited." In: *IEEE Sensors Journal* 17.11 (2017), pp. 3257–3268.

[140] Haythem Rehouma, Rita Noumeir, Sandrine Essouri, and Philippe Jouvet. "Advancements in Methods and Camera-Based Sensors for the Quantification of Respiration." In: *Sensors* 20.24 (2020), p. 7252. ISSN: 1424-8220. DOI: 10.3390/s20247252. URL: https://www.mdpi.com/1424-8220/20/24/7252.

[141] Daniela Ridel, Eike Rehder, Martin Lauer, Christoph Stiller, and Denis Wolf. "A literature review on the prediction of pedestrian behavior in urban scenarios." In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 3105–3112.

[142] Mary Roberts, David Mongeon, and Francois Prince. "Biomechanical parameters for gait analysis: a systematic review of healthy human gait." In: *Phys. Ther. Rehabil* 4.6 (2017).

[143] Daniel Roetenberg, Henk Luinge, and Per Slycke. "Xsens MVN: full 6DOF human motion tracking using miniature inertial sensors." In: *Xsens M T BV, Tech. Rep* (2009).

[144] Chiara Romano, Emiliano Schena, Sergio Silvestri, and Carlo Massaroni. "Non-Contact Respiratory Monitoring Using an RGB Camera for Real-World Applications." In: *Sensors (Basel, Switzerland)* 21.15 (2021). DOI: 10.3390/s21155126. (Visited on 04/23/2022).

[145] Christian Schaller, Jochen Penne, and Joachim Hornegger. "Time-of-flight sensor for respiratory motion gating." In: *Medical Physics* 35.7Part1 (2008), pp. 3090–3093.

[146] Martin Schätz, Fabio Centonze, Jiří Kuchyňka, Ondřej Ťupa, Oldřich Vyšata, Oana Geman, and Aleš Procházka. "Statistical recognition of breathing by MS Kinect depth sensor." In: *Computational Intelligence for Multimedia Understanding (IWCIM), 2015 International Workshop on*. IEEE. 2015, pp. 1–4.

[147] Martin Schätz, Aleš Procházka, Jiří Kuchyňka, and Oldřich Vyšata. "Sleep Apnea Detection with Polysomnography and Depth Sensors." In: *Sensors (Basel, Switzerland)* 20.5 (2020). DOI: 10.3390/s20051360. (Visited on 07/14/2022).

[148] Osamu Shigeta, Shingo Kagami, and Koichi Hashimoto. "Identifying a moving object with an accelerometer in a camera view." In: *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2008, pp. 3872–3877.

[149] Yiu Cheung Shiu and Shaheen Ahmad. "Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX= XB." In: (1987).

[150] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. "Real-time Human Pose Recognition in Parts from Single Depth Images." In: *Commun. ACM* 56.1 (Jan. 2013), pp. 116–124. ISSN: 0001-0782.

[151] Aneesha Singh, Stefano Piana, Davide Pollarolo, Gualtiero Volpe, Giovanna Varni, Ana Tajadura-Jiménez, Amanda CdeC Williams, Antonio Camurri, and Nadia Bianchi-Berthouze. "Go-with-the-flow: tracking, analysis and sonification of movement and breathing to build confidence in activity despite chronic pain." In: *Human–Computer Interaction* 31.3-4 (2016), pp. 335–383.

[152] V. Soleimani, M. Mirmehdi, D. Damen, J. Dodd, S. Hannuna, C. Sharp, M. Camplani, and J. Viner. "Remote, Depth-Based Lung Function Assessment." In: *IEEE Transactions on Biomedical Engineering* 64.8 (Aug. 2017), pp. 1943–1958. ISSN: 0018-9294. DOI: 10.1109/TBME.2016.2618918.

[153] Olivier Steichen, Gilles Grateau, and Eric Bouvard. "Respiratory rate: the neglected vital sign." In: *The Medical journal of Australia* 189.9 (2008), pp. 531–532. ISSN: 0025-729X. DOI: 10.5694/j.1326-5377.2008.tb02164.x. (Visited on 04/21/2022).

[154] Sebastian Stein and Stephen J McKenna. "Accelerometer localization in the view of a stationary camera." In: *2012 Ninth Conference on Computer and Robot Vision*. IEEE. 2012, pp. 109–116.

[155] Thomas Stiefmeier, Daniel Roggen, and Gerhard Tröster. "Gestures are strings: efficient online gesture spotting and classification using string matching." In: *Proceedings of the ICST 2nd international conference on Body area networks*. 2007, pp. 1–8.

[156] K Song Tan, Reza Saatchi, Heather Elphick, and Derek Burke. "Real-time vision based respiration monitoring system." In: *7th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*. IEEE. 2010, pp. 770–774.

[157] The Free Software Foundation. *GCC, the GNU Compiler Collection*, *http://gcc.gnu.org*. Version 10.1.0.

[158] Yushuang Tian, Xiaoli Meng, Dapeng Tao, Dongquan Liu, and Chen Feng. "Upper limb motion tracking with the integration of IMU and Kinect." In: *Neurocomputing* 159 (2015), pp. 207–218.

[159] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. "Total capture: 3d human pose estimation fusing video and inertial sensors." In: *Proceedings of 28th British Machine Vision Conference*. 2017, pp. 1–13.

[160] Andrea Valenzuela, Nicolás Sibuet, Gemma Hornero, and Oscar Casas. "Non-Contact Video-Based Assessment of the Respiratory Function Using a RGB-D Camera." In: *Sensors (Basel, Switzerland)* 21.16 (2021). DOI: 10.3390/s21165605. (Visited on 07/14/2022).

[161] Kristof Van Laerhoven, Eugen Berlin, and Bernt Schiele. "Enabling efficient time series analysis for wearable activity data." In: *2009 International Conference on Machine Learning and Applications*. IEEE. 2009, pp. 392–397.

[162] Federico Vicentini. "Collaborative robotics: a survey." In: *Journal of Mechanical Design* 143.4 (2021), p. 040802.

[163] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. "Recovering accurate 3d human pose in the wild using imus and a moving camera." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 601–617.

[164] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus." In: *Computer graphics forum*. Vol. 36. 2. Wiley Online Library. 2017, pp. 349–360.

[165] Hao Wang, Daqing Zhang, Junyi Ma, Yasha Wang, Yuxiang Wang, Dan Wu, Tao Gu, and Bing Xie. "Human Respiration Detection with Commodity Wifi Devices: Do User Location and Body Orientation Matter?" In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, 25–36. ISBN: 9781450344616. DOI: 10.1145/2971648.2971744. URL: https://doi.org/10.1145/2971648.2971744.

[166] Xuyu Wang, Chao Yang, and Shiwen Mao. "TensorBeat: Tensor Decomposition for Monitoring Multiperson Breathing Beats with Commodity WiFi." In: *ACM Trans. Intell. Syst. Technol.* 9.1 (Sept. 2017), 8:1–8:27. ISSN: 2157-6904. DOI: 10.1145/3078855. URL: http://doi.acm.org/10.1145/3078855.

[167] Jakob Wasza, Sebastian Bauer, Sven Haase, and Joachim Hornegger. "Sparse principal axes statistical surface deformation models for respiration analysis and classification." In: *Bildverarbeitung für die Medizin 2012*. Springer, 2012, pp. 316–321.

[168] Jakob Wasza, Peter Fischer, Heike Leutheuser, Tobias Oefner, Christoph Bert, Andreas Maier, and Joachim Hornegger. "Real-time respiratory motion analysis using 4-D shape priors." In: *IEEE Transactions on Biomedical Engineering* 63.3 (2016), pp. 485–495.

[169] Dirk Weenk, Bert-Jan F Van Beijnum, Chris TM Baten, Hermie J Hermens, and Peter H Veltink. "Automatic identification of inertial sensor placement on human body segments during walking." In: *Journal of neuroengineering and rehabilitation* 10.1 (2013), p. 31.

[170] Juyang Weng, Paul Cohen, Marc Herniou, et al. "Camera calibration with distortion models and accuracy evaluation." In: *IEEE Transactions on pattern analysis and machine intelligence* 14.10 (1992), pp. 965–980.

[171] Udaya Wijenayake and Soon-Yong Park. "Real-Time External Respiratory Motion Measuring Technique Using an RGB-D Camera and Principal Component Analysis." In: *Sensors* 17.8 (2017), p. 1840.

[172] X. Wu, Y. Wang, C. Chien, and G. Pottie. "Self-calibration of sensor misplacement based on motion signatures." In: *2013 IEEE International Conference on Body Sensor Networks*. May 2013, pp. 1–5. DOI: 10.1109/BSN.2013.6575504.

[173] Junyi Xia and R Alfredo Siochi. "A real-time respiratory motion monitoring system using KINECT: Proof of concept." In: *Medical physics* 39.5 (2012), pp. 2682–2685.

[174] A. D. Young, M. J. Ling, and D. K. Arvind. "Distributed estimation of linear acceleration for improved accuracy in wireless inertial motion capture." In: *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks - IPSN '10*. ACM Press, 2010.

[175] Muhamad Nurul Hisyam Yunus, Mohd Hafiidz Jaafar, Ahmad Sufril Azlan Mohamed, Nur Zaidi Azraai, and Md Sohrab Hossain. "Implementation of kinetic and kinematic variables in ergonomic risk assessment using motion capture simulation: A review." In: *International Journal of Environmental Research and Public Health* 18.16 (2021), p. 8342.

[176] Youwei Zeng, Dan Wu, Ruiyang Gao, Tao Gu, and Daqing Zhang. "FullBreathe: Full Human Respiration Detection Exploiting Complementarity of CSI Phase and Amplitude of WiFi Signals." In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.3 (Sept. 2018). DOI: 10.1145/3264958. URL: https://doi.org/10.1145/3264958.

[177] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. "Fusing wearable imus with multi-view images for human pose estimation: A geometric approach." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2200–2209.

[178] Yang Zheng, Ka-Chun Chan, and Charlie C. L. Wang. "Pedalvatar: An IMU-based real-time body motion capture system using foot rooted kinematic model." In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, September 14-18, 2014*. 2014, pp. 4130–4135.

[179] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. "Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 384–400.

[180] Tobias Zimmermann, Bertram Taetz, and Gabriele Bleser. "IMU-to-segment assignment and orientation alignment for the lower body using deep learning." In: *Sensors* 18.1 (2018), p. 302.

[181] Andreas Zinnen, Kristof Van Laerhoven, and Bernt Schiele. "Toward Recognition of Short and Non-repetitive Activities from Wearable Sensors." In: *Ambient Intelligence: European Conference, AmI 2007, Darmstadt, Germany, November 7-10, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 142–158.