Bachelor Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

# Evaluation framework for infant 3D pose extraction from RGB images using RGB-D cameras and motion capture system

Noemi Vaculínová

Supervisor: doc. Mgr. Matěj Hoffmann, Ph.D.
Supervisor–specialist: Valentin Marcel, Ph.D
Study program: Cybernetics and Robotics
May 2023

# BACHELOR‘S THESIS ASSIGNMENT

## I. Personal and study details

Student's name: **Vaculínová Noemi** Personal ID number: **498821**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Cybernetics and Robotics**

## II. Bachelor’s thesis details

Bachelor’s thesis title in English:

**Evaluation Framework for Infant 3D Pose Extraction from RGB Images Using RGB-D Cameras and Motion Capture System**

Bachelor’s thesis title in Czech:

**Vyhodnocení přesnosti extrakce 3D pozice kojenců z RGB obrázků pomocí RGB-D kamer a systému sledování pohybu**

Guidelines:

Automatic extraction of infant 3D poses from RGB images or videos only is important in developmental and clinical psychology. The movement patterns extracted can be used for automatic diagnosis of psychomotor disabilities such as cerebral palsy, without the need for bringing infants physically to the therapist. State-of-the art methods for adults (e.g. Simplify_x [1]) are trained on datasets with motion capture data as ground truth. Such data is hard to obtain for young children. An existing dataset [2] is based on RGB-D data of moving infants; the 3D shape and pose extraction method (SMIL) is described in [3]. No infant dataset with motion capture and RGB-D exists. This thesis will fill this gap by creating a small dataset (few infants) and developing a corresponding evaluation method.

Instructions:

1. Get familiar with the Qualisys motion capture system, marker placements on infants and adults[4].
2. Get familiar with RGB-D cameras and point clouds processing methods such as segmentation of the background.
3. Prepare the data acquisition, including a suit with markers attached that the infants can wear.
4. Collect the data - minimum 2 infants.
5. Develop a method for temporal synchronization of the data collected.
6. Perform alignment between 3D motion capture keypoints and RGB-D point cloud.
7. Perform point-sets registration from SMIL’s mesh to the captured RGB-D point cloud.
8. Propose a criterion to evaluate the kinematics dissimilarities between 3D keypoints based on state-of-the-art criteria [5]: mean joint position error, mean joint angle error, etc.

Bibliography / sources:

[1] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., & Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10975-10985).

[2] Hesse, N. et al.. (2019). Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set. Lecture Notes In Computer Science, 32-49.

[3] Hesse, N., Pujades, S., Black, M. J., Arens, M., Hofmann, U. G., & Schroeder, A. S. (2020). Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. IEEE transactions on pattern analysis and machine intelligence, 42(10), 2540-2551.

[4] Wu, G. et al. (2005). ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion--Part I & Part II. Journal of biomechanics.

[5] Wang, J. et al. (2021) "Deep 3D human pose estimation: A review", Computer Vision and Image Understanding, 210, p. 103225.

Name and workplace of bachelor’s thesis supervisor:

**doc. Mgr. Matěj Hoffmann, Ph.D. Vision for Robotics and Autonomous Systems FEE**

Name and workplace of second bachelor’s thesis supervisor or consultant:

**Valentin Marcel, Ph.D. Vision for Robotics and Autonomous Systems FEE**

Date of bachelor’s thesis assignment: **12.01.2023** Deadline for bachelor thesis submission: **26.05.2023**

Assignment valid until: **22.09.2024**

|  |  |  |
|---|---|---|
| doc. Mgr. Matěj Hoffmann, Ph.D.<br>Supervisor’s signature | prof. Ing. Tomáš Svoboda, Ph.D.<br>Head of department’s signature | prof. Mgr. Petr Páta, Ph.D.<br>Dean’s signature |

## III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt

Student’s signature

# Acknowledgements

First, I would like to thank my supervisor, Matj Hoffmann, for valuable insights and always maintaining a positive and calm approach in every situation.

I am also very grateful for the valuable guidance and support in experiments and data processing from Valentin Marcel, my second supervisor.

Furthermore, I must express my gratitude to Sergiu Tcaci Popescu for insight into the field of psychology and the tremendous support and kind words he had for me, while I was writing these pages.

I would also like to acknowledge Filipe Gama for processing the video footages and through proofreading of this text, and Jason Khoury for help in experiments and the invaluable support in labeling motion capture data.

Furthermore, I am grateful for the parents who participated in this study and contributed to the results.

I would like to express my gratitude and appreciation to Petr Beránek for bringing the first doll for the experiments and building the overhead camera rig.

My deepest gratitude belongs to my family and friends, who supported me through this journey. Mainly to my father, who read the first drafts of this thesis and had only kind words.

Lastly, I extend my thanks to anyone I did not mention here, as this space is too small to express all the gratitude I have and name everyone, without whom this work would not have happened.

# Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských prací.

V Praze dne 26. května 2023

......................
Noemi Vaculínová

# Abstract

Infant motion analysis is important in diagnosing motor and cognitive developmental disorders such as cerebral palsy. Automated extraction of infant movement from only RGB videos could facilitate early and remote diagnostics of such disorders prior to an in-person assessment by a trained professional. Human pose estimation techniques from images and videos with outputs in 2D and even 3D are available. Fewer methods are available for infants. However, the accuracy of these methods is not known and datasets with corresponding ground truth data from motion capture systems are missing. This work bridges this gap by providing recordings of infants that were taken simultaneously with motion capture, RGB, and RGB-D cameras. Different methods of marker placement, synchronization, and alignment of different data streams, and comparison with pose estimation from RGB only are presented. Two infants (3 and 8 months old) were recorded. After completing the analysis, anonymizing the data, and adding more recordings, we plan to release this as a public dataset for the community.

**Keywords:** infant motion extraction, human pose estimation, RGB videos, motion capture system, RGB-D cameras, motion extraction validation, human keypoint extraction

**Supervisor:** doc. Mgr. Matěj Hoffmann, Ph.D.

# Abstrakt

Analýza pohybu kojenců je důležitá při diagnostice motorických a kognitivních vývojových poruch, jako je například dětská mozková obrna. Automatizovaná extrakce pohybu kojenců z RGB videí by mohla usnadnit včasnou a vzdálenou diagnostiku těchto poruch před osobním posouzením vyškoleným odborníkem. K dispozici jsou techniky odhadu polohy člověka ze snímků a videí s výstupy ve 2D a dokonce i ve 3D. Pro kojence však existuje málo metod, Jejichž přesnost však není známa a chybí datasety s odpovídajícími ground truth údaji z motion capture systémů. Tato práce výše zmíněnou mezeru zaplňuje tím, že poskytuje záznamy kojenců, které byly pořízeny současně kamerami RGB a RGB-D a motion capture systémem. Jsou zkoumány různé metody umístění značek, synchronizace a zarovnání různých datových toků a srovnání s odhadem polohy pouze z RGB dat. Byly pořízeny záznamy dvou kojenců (ve věku 3 a 8 měsíců), které po dokončení analýzy, anonymizace dat a přidání dalších nahrávek je plánováno zpřístupnit jako veřejný dataset pro komunitu.

**Klíčová slova:** extrakce pohybu kojenců, odhad polohy člověka, videa RGB, motion capture systém, kamery RGB-D, validace extrakce pohybu, extrakce klíčových bodů člověka

**Překlad názvu:** Vyhodnocení přesnosti extrakce 3D pozice kojenců z RGB obrázků pomocí RGB-D kamer
a systému sledování pohybu

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation

Researchers in various fields are interested in the study of infants' movements. For example, in developmental psychology, infant movements are studied to understand the time course of motor development and the factors that affect it, or to infer from motor behavior hidden variables such as goal-directedness and intentionality [1, 2]. In healthcare, infant movements are used to evaluate the time course of developmental milestones of motor abilities. These milestones encompass the ability to turn the head, remain in a seated position without external support, crawl, grasp, and locomote. Furthermore, analysis of infant movement helps to detect early developmental disorders and pathologies, including cerebral palsy [3]. In robotics, particularly in developmental robotics, infant movement is used as an inspiration for learning motor strategies and motor control, but also to simulate other inputs (proprioception, muscular activation, vision that are covert) resulting from an overt movement [4].

The data necessary for the study of infant movements can be captured using various methods, ranging from the use of motion capture platforms, which are currently the gold standard in terms of data accuracy, to the use of human manual coding (i.e. labeling) of video recordings, which is very time consuming. The use of a motion capture platform, despite being a gold standard, has several limitations that make the search for alternative solutions necessary. For instance, motion capture platforms are almost exclusively used in laboratory settings and can hardly be used at the infants' homes; a motion capture platform also requires a considerable financial investment and specific expertise both to operate it and to extract kinematic data from the raw data.

On the other hand, video recording devices such as smartphones are rather widely available and cost but a tiny fraction of motion capture platform costs (the costs differ by at least two orders of magnitude); also, they do not require any specific expertise to operate and can be flexibly used in all places to capture infant's spontaneous behavior frequently and at home, in different light settings, and at varying distances. Having the ability to extract motion data from video recordings made with smartphones or other video-recording devices, while maintaining an acceptable level of accuracy, would thus represent a significant contribution to the community of scientists and specialists who use infant motion data, i.e. for early diagnostics of developmental disorders. It should also be mentioned that video recording devices that can also record depth information, referred to as RGB-D cameras, offer an intermediate solution and in recent years these devices have become accessible at a very moderate cost, similar to that of smartphones.

The general question of this thesis is how can such alternative solutions to movement extraction be proposed while ensuring data quality that would be appropriate for application in psychology, health-sciences, and robotics. The Humanoid and Cognitive Robotics Team at the Czech Technical University in Prague, using open-source software tools, proposed a methodology and has established a processing pipeline to extract infant motion kinematics directly from simple RGB video recordings and showed how it could be used to study infant spontaneous behaviors [5]. However, the kinematic results provided by using this methodology and pipeline have not been tested against the gold standard of motion capture as there is no infant dataset with ground truth movement.

## 1.2 Goals

The primary goal of this thesis is to capture video recordings of two infants in the supine position and generate the corresponding joint position data and validate the SMIL pipeline. Motion capture technology, known for its precision, will serve in combination with RGB-D data as the ground truth dataset against which any method of joint estimation can be compared using selected evaluation metrics.

## 1.3 Structure of thesis

The thesis is structured as follows. First, an overview of state-of-the-art methods for extracting movement from videos, estimating joints from motion capture, and integrating information from motion capture and RGB-D cameras is provided in Chapter 2. This is followed by an explanation of the methodologies employed to acquire ground-truth data using optical motion capture and RGB-D cameras in Chapter 3. Given the utilization of multiple data sources, synchronization techniques are discussed as a crucial aspect of the investigation. Evaluation methods and metrics are also explored, as one of the thesis objectives involves comparing motion capture joints with the outcomes derived from the pipeline of the Humanoid and Cognitive Robotics Team.

Subsequently, in Chapter 4 detailed discussions on all experiments carried out that led to the infant recordings are presented, with particular emphasis on determining the optimal setup and solution for infant ground truth acquisition. Chapter 5 provides a discussion and conclusion.

# Chapter 2

# Related Work

In this chapter, an overview of 3D pose extraction methods, body models, evaluation methods and metrics is presented.

## 2.1 Overview of 3D pose extraction from RGB images

Motion capture with physical markers is difficult and sometimes not even possible in early infancy or with preterm infants. The following question arises: How to evaluate motion otherwise? This is why some studies evaluated alternative methods of pose extraction from RGB and RGB-D images. The four main lines of this research are: direct pose estimation, lifting from 2D to 3D, model-based pose estimation, and infant 3D pose extraction from RGB images. The last method – infant 3D pose extraction from RGB images – is of special interest for both the research on developmental robotics of our team and for the applications in early infancy. The following paragraphs provide some detail on each of these methodologies and the specific problems with infant pose estimation.

Direct pose estimation relies on the estimation of joint positions directly from the image. Pavlakos et al. [6] represent the joints with volumetric probability heatmaps. They train a convolutional neural network with coarse to fine supervision to locate the joints. Other works employ regression models to generate and process the heatmaps [7, 8]. MMPose [9] is an open source toolbox for pose estimation based on deep learning.

Many works have been inspired by the development of 2D human pose estimation algorithms and aim to use 2D pose estimation results for 3D human pose estimation. Different methods of this 2D to 3D lifting have been proposed to fuse 2D joint heatmaps with 3D image cues [10], such as using neural networks [11], long-short-term memory (LSTM) [12], and generative adversarial networks [13].

Model-based pose estimation is based on fitting a learned body model to images or video footage. One of the commonly used body models is the *Skinned Multi-Person Linear Model* (SMPL) [14], which uses high-resolution 3D scans of subjects in a wide variety of poses from the CAESAR dataset [15] to create realistic body models, with variety in shape and gender. The SMPL model has a male, female, and gender neutral version to best accommodate all bodies. *SMPL eXpressive* (SMPL-X) and SMPLify-X models [16] are extensions of SMPL with added articulated hands and facial expressions. The SMPL-X model uses the FLAME head model and the MANO hand model for expression and hand articulation. While model-based, SMPL and SMPLify-X still require a first step of 2D pose estimation

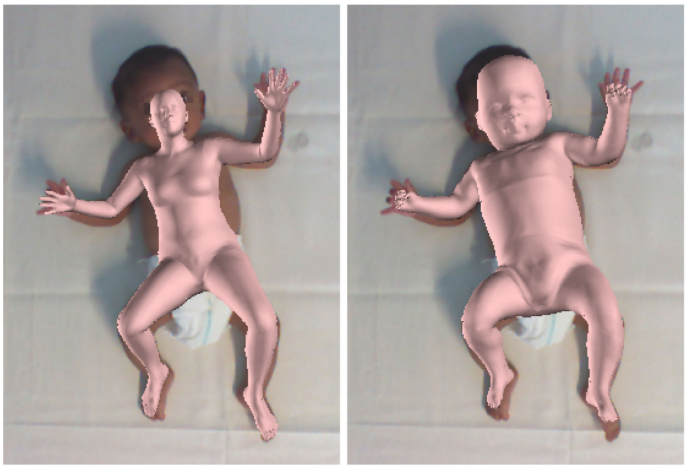


**Figure 2.1:** (left) Simply scaling the SMPL adult body model and fitting it to an infant does not work as body proportions significantly differ. (right) The proposed SMIL model properly captures the infants' shape and pose. Figure and caption taken from [19].

and use it to optimize and propose a 3D model.

RGB-D cameras are used in studies with preterm infants [17], with neonates in intensive care [18] and for infant motion analysis [19]. However, the information provided by RGB-D has limits: precise kinematics of body parts are hard to compute; if one needs to estimate more precisely the kinematics of different body parts (e.g. hand, legs), one needs a method that would allow to estimate precise spatial positions and joint angles of infant body. This is why in this work the RGB-D data is combined with data from optical motion capture.

Using an adult model (such as SMPL or SMPLify-X) for the estimation of infant pose is not sufficient, because body proportions are vastly different, as illustrated in Fig. 2.1. Hesse et al. modified the SMPL model [14] to fit the proportions of the infant and created the *Skinned Multi-Infant Linear body model* (SMIL) [19, 20]. They created the model using many low-quality RGB-D recordings of freely moving infants [21]. Using this model and RGB-D recordings of infant spontaneous movement Hesse et al. created the *Moving INfants In RGB-D* (MINI-RGBD) dataset [22]. The movement of the infant is tracked and projected on the SMIL model with realistic shapes and textures. The dataset consists of RGB and depth footage of the synthetic infants with ground-truth information. This thesis proposes the acquisition of real infant videos, as it is more realistic than a 3D model. Keypoints are points, that are crutial for the 3D pose estimation, they are mainly joints, however, eyes, ears and any other parts of the hunam body can be keypoints. Joint positions refer to the position of joint centers in the context of 3D pose estimation.

## 2.2 Overview of methods for determining joint positions

Estimation of joint centers without the ability to look inside the joints, e.g. with X-rays, is a challenging subject to tackle in human gait and posture analysis. This section gives an overview of some methods of joint position estimation using markers with optical motion capture systems.

### 2.2.1 Anatomical landmarks

Markers on anatomical landmarks [23, 24] are used to locate joint centers. This method has many pitfalls. There is movement of tissue and skin between the bony landmarks and the marker, resulting in the marker potentially moving. Furthermore, perfectly palpating these landmarks, as the method requires, is difficult even for experienced physicians, as van Sin Jan states in [25]. As a result, the obtained measurements exhibit an accuracy level of approximately 2 cm [26]. This may not seem as much, but for bodies of small dimensions, such as those of neonates or infants, this margin of error represents a high proportion. A solution to reduce it would be required for precise applications, such as medical care. In addition, the complexity of palpating the landmarks and placing markers in the correct positions increases in a moving and non-cooperative infant.

### 2.2.2 Sphere fitting

Ehrig et al. in a survey [27] report on sphere-fitting methods that assume that the center of rotation (CoR) of the observed joint is stationary. It takes advantage of the fact that the radius is constant. CoR is calculated by minimizing the Euclidean distances between the sphere and marker positions [28]. According to Ehrig et al., this method may not provide the highest level of accuracy, as it relies on the assumption that some joints remain stationary.

A different approach to sphere fitting are the transformation techniques. In this case, rigid body transformations are used to find the CoR [29]. One segment is made stationary using the transformations, whereas the moving segment is fitted to a sphere. The center of this sphere is the desired CoR. This method was chosen because Ehrig et al. [27] suggest that its reliability and precision is greater than the method in the previous paragraph.

## 2.3 Background of 3D pose extraction from optical motion capture

Optical motion capture is regarded as a source of accurate gold standard information. It is used for motion tracking and analysis. For example, Herda et al. [30] use passive marker-based optical motion capture for adult motion reconstruction. Kanazawa et al. [31] focus on infant motion capture. They combine the information from an optical motion capture system with bone and muscle simulations in OpenSim [32] to detect the joint centers. Motion capture is also sometimes combined with the information from RGB-D cameras. For example, Zhao et al. [33] use this combination for hand movement and position capture.

## 2.4 Evaluation methods and metrics

One part of this thesis focuses on the evaluation of the infant movement extraction pipeline as discussed in Sec. 1.1. The evaluation of pose estimation methods is mostly calculated using a ground-truth model. In most cases, as written in [34], the Mean per joint position error (*MPJPE*) is used [35], which compares the joint positions of the ground truth joints and the estimated joints. Van Crombrugge et al. [36] propose a metric that computes the joint angle error. Joint angles are crucial to evaluate motion and are not affected by the scale and alignment of ground truth and estimated data. It has also been used in infant motion extraction [37].

Evaluation without relying on ground truth data is done by calculating the standard deviation of bone length. Bone length is the distance between two joints connected by a bone in the human body. All bones have a constant length, so any extension or shortening is a sign of inaccuracy.

## 2.5 Thesis contribution

The literature study indicated missing dataset with real infants and ground truth using optical motion capture. This work proposes a method for acquisition of ground truth of infant spontaneous movement in laboratory conditions using motion capture and RGB-D cameras. This ground truth data should provide validation of the motion extraction from only RGB or RGB-D videos.

# Chapter 3

# Materials and Methods

## 3.1 Equipment

To capture the movements of the infant, an optical motion capture system that tracks sensors on specific body parts was employed in this work. The captured motion data was then integrated with the data obtained from RGB-D cameras, which simultaneously record depth and color information. This section discusses the mechanics and setups of these cameras related to our experiments.

### 3.1.1 Optical motion capture system with passive markers

Optical motion capture relies on triangulating the positions of retro-reflective markers within a Cartesian coordinate system. To achieve this, multiple infrared cameras are strategically positioned around a specific area of interest to track retro-reflective markers. Retro-reflective refers to the property of a material or surface to reflect light back towards its source.

Surrounding the light capturing sensor of the infra red camera, a circular arrangement of infrared light-emitting diodes (LEDs) is installed. These LEDs emit infrared light that reflects off the markers, allowing the cameras to capture their positions in a two-dimensional image. The triangulation process involves using the known relative positions and orientations of the cameras and the captured images. From this information, the x, y, and z coordinates of each marker can be precisely determined. Figure 3.1 provides a visual representation of the mechanics involved in this process. The camera layout used in the experiments is described in Sec. 3.1.1.

#### Retro-reflective markers

Spherical markers coated in retro-reflective paint and placed on stands are commonly used in conjunction with infrared motion capture cameras. In certain instances, half-spheres [31] or reflective tape cut into various shapes are used as alternative marker options. The elevated spherical markers offer optimal visibility to cameras from multiple angles, reducing the likelihood of occlusion caused by body limbs.

Typically, spherical markers are affixed to body parts using double-sided tape or incorporated into specialized suits with velcro attachments. However, these suits are not suitable for infants, necessitating the exploration of alternative attachment methods. The number

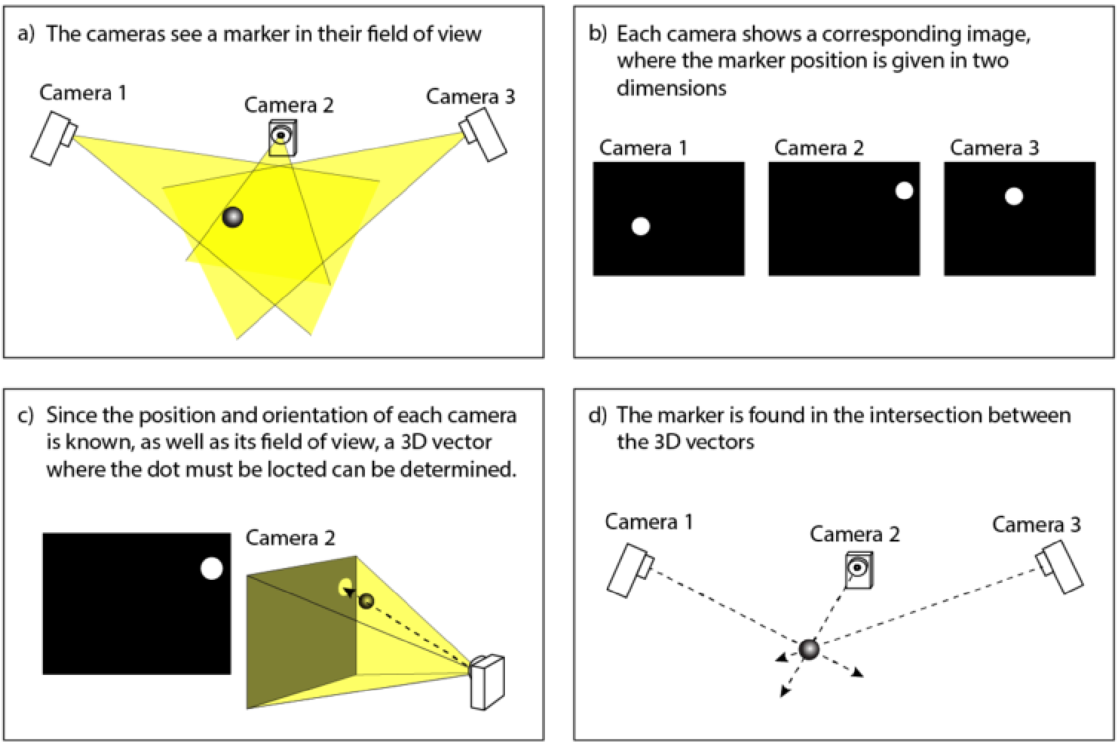

a) The cameras see a marker in their field of view
Camera 1
Camera 2
Camera 3
b) Each camera shows a corresponding image, where the marker position is given in two dimensions
Camera 1
Camera 2
Camera 3
c) Since the position and orientation of each camera is known, as well as its field of view, a 3D vector where the dot must be locted can be determined.
Camera 2
d) The marker is found in the intersection between the 3D vectors
Camera 1
Camera 2
Camera 3

**Figure 3.1:** Through the integration of images captured by individual infrared motion capture cameras, a three-dimensional representation of each marker is generated. Figure taken from [38].

and specific placement of markers vary depending on the chosen data processing method. In human motion analysis, every main joint of the body is tracked, meaning that the minimal marker layout is a marker on each joint. However, the information from this set of markers is limited and is not sufficient for proper motion analysis. Using more markers provides more information, leading to higher precision. On the other hand, too many markers lead to marker merging and discomfort of the infant. An optimal compromise must be found in Chapter 4.

In our experimental setup, spherical markers of three different diameters and shapes cut from retro-reflexive tape, varied marker layouts, and alternative adhesion methods were utilized. Detailed discussions on these aspects can be found in the corresponding sections related to the experiments (see Chapter 4).

### Calibration

Before each motion capture session, it is necessary to calibrate the volume of interest. The cameras are calibrated with respect to an origin point, which is denoted by an L-frame positioned within the volume of interest. The calibration process is started in the motion capture recording software QTM (Qualisys Track Manager) and a calibrated wand must be moved through the volume of interest. This process enables the determination of the

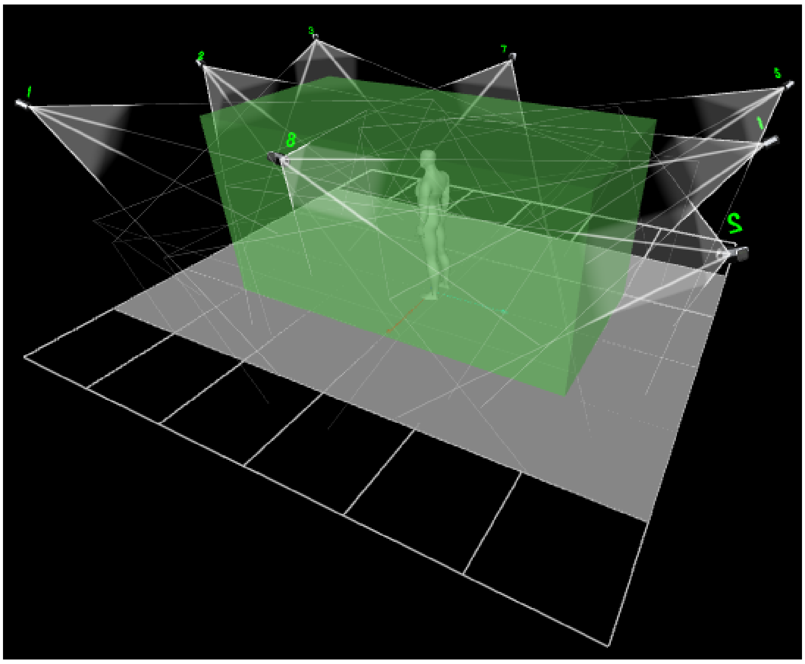


**Figure 3.2:** The layout of the eight infra red motion capture cameras. Figure taken from [39].

**Figure 3.3:** A panoramatic photo of the infa red motion capture camera setup in our laboratory. Photo taken by Adadm Rojík.

relative positions of the cameras.

### Camera setup and parameters

The Qualisys motion capture system in our laboratory consists of eight Miqus M3 infrared cameras supplemented by a Miqus video camera. The video camera provides RGB video synchronized with the motion capture data.

The layout of the cameras is as follows. Infrared cameras are mounted on the walls of the laboratory sized 6.2 by 6.2 meters, approximately 3 meters from the ground. One camera is in each corner and in the middle of every wall. They are pointed at the middle of the room as shown in Fig. 3.2. The video camera can be mounted on a tripod anywhere in the room. The position of the Miqus video camera changed through the experiments and is always disclosed in the relevant experiments in Chapter 4.

QTM enables customizing of the camera parameters for each camera individually. In our experiments, two sets of parameters were used, one for the video and one for the eight infra-red Miqus M3 cameras.

The video camera recorded frames sized 1088x1472 px at 25 fps. Auto-balancing and auto-exposure of the images could be turned on or off. The infrared motion capture cameras captured images sized 1823x1087 px at a sampling rate of up to 650 fps. To optimize the visibility of markers, the exposure time and marker threshold can be changed.

Most of the parameters were changed through experiments because they depend, among other things, on the lighting conditions. These changes are disclosed in the corresponding sections of Chapter 4.

### 3.1.2 RGB-D camera

RGB-D camera is equipped with a depth sensor and a RGB sensor, allowing simultaneous depth perception and color imaging. Color is detected and recorded using a regular RGB camera. Depth information is obtained by calculating the time of flight (ToF) of an invisible infrared pulse of light. This results in a depth map. From these data, the camera generates a colored point cloud, which can be processed, e.g. in Matlab.

#### Camera setup and parameters

Two different models of RGB-cameras were used in the experiments. First, the Intel RealSense D435 was employed, with a depth sensor offering a depth range of up to three meters, providing accurate depth measurements with a resolution of up to 1280x720. The D435 was later replaced by the newer Intel RealSense D455 camera with a depth range of up to six meters and a resolution of up to 1280x720. Both cameras can record RGB data at 90 fps and depth data at up to 100 fps. These parameters are the maxima provided by the manufacturer; however, higher resolution and framerate are applied in expense of the camera field of view and the file size. Therefore, the framerate and resolution were optimized for each experiment specifically (see Chapter 4).

Depth measurements are limited by occlusions resulting in missing depth information on certain parts of the body. To reduce this limitation, two cameras were used to capture the movement, as they had two angles, so that the information could be supplemented. The synchronization of pointclouds is explained in Sec. 3.2.3.

#### RGBD and Lighting condition

As the RGB-D camera uses reflections and ToF to measure depth data, problems occur when recording highly reflective and non-reflective materials. A derivation of this problem had to be solved in our experiments. Bright sunlight and overhead lighting created luminous spots on the floor and even on the infant clothing. These luminous spots created gaps in the pointclouds. Therefore, the overhead lighting was turned off and the blinds were shut throughout the experiments.

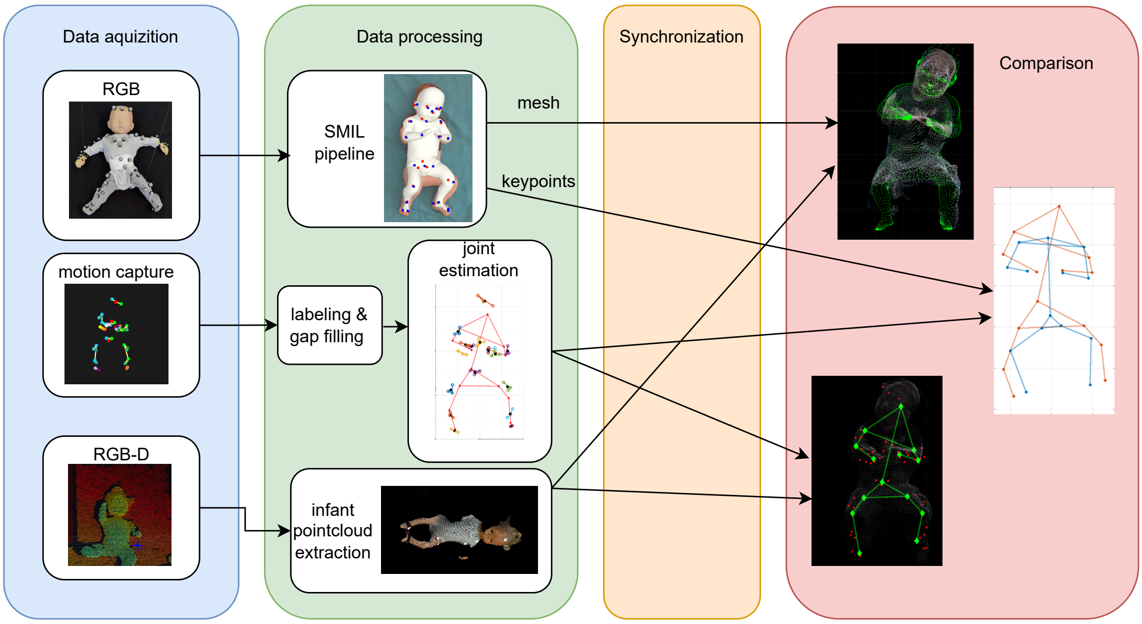


**Figure 3.4:** The data processing of all three streams of information.

## 3.2 Data processing

In order to compare the ground truth to the results of SMIL algorithm, multiple steps have to be performed after recording each data stream. All data streams are not directly exploitable after recording. First, all data has to be synchronized together. Marker positions from motion capture data can be missing, and RGB data capture the full room and is very noisy. It has to be cropped and filtered. The following sections describe this data processing as pictured in Fig. 3.4.

### 3.2.1 Mocap postprocessing

In this section, the processing steps for the data obtained through optical motion capture from the final infant experiment (4.2) are outlined. The process involved labeling and filling trajectories within the Qualisys Track Manager (QTM) software. Rigid body approximations were applied to each body part, followed by the extraction of joint centers from the calculated centers of rotation within the rigid bodies.

#### Labeling and gap filling

From the Qualisys motion capture system point trajectories were generated (see Sec. 3.1.1). These trajectories required manual labeling to assign the markers to the corresponding body parts. The labeling process was time-consuming due to occlusions and merging of markers when they closely overlapped.

Even after manual labeling, certain gaps remained within the trajectories. In the case of the forearms, which had six markers each, a rigid body approximation was possible when three or more markers were still visible. However, in order to minimize the number of markers, other body parts lacked redundancy, making it challenging to fill gaps. Polynomial curve regression was used in QTM to fill gaps up to ten frames long (100 ms), while larger gaps were excluded from the dataset.

### Creating rigid bodies

A rigid body is an idealized object or system in which the distances between its particles remain constant. The position and orientation of a point on a rigid body can be determined by a rotation matrix and a translation vector. To define a rigid body in a three-dimensional space, the positions of at least three points on the object must be known [40].

The markers are fixed together in clusters on patches inspired by Kanazawa et. al [31] (see Sec. 4.1.7), which minimizes the positional deviation relative to the cluster. Also, except for hand and foot markers, all clusters consist of three individual markers, which is the minimum, to describe a rigid body as stated above.

Each cluster c corresponding to a rigid body was represented in the initial coordinate system by the position of the central point of the cluster $\boldsymbol{t}_{\text{fc}}$ and a rotation matrix $\boldsymbol{R}_{\text{fc}}$. The points in clusters were represented by their Carthesian coordinates:

$$\boldsymbol{p}_{\text{fc1}} = \begin{bmatrix} x_{\text{fc1}} \\ y_{\text{fc1}} \\ z_{\text{fc1}} \end{bmatrix}, \ \boldsymbol{p}_{\text{fc2}} = \begin{bmatrix} x_{\text{fc2}} \\ y_{\text{fc2}} \\ z_{\text{fc2}} \end{bmatrix}, \ \boldsymbol{p}_{\text{fc3}} = \begin{bmatrix} x_{\text{fc3}} \\ y_{\text{fc3}} \\ z_{\text{fc3}} \end{bmatrix}, \tag{3.1}$$

where $\boldsymbol{p}_{\text{fc1}}$ is the position of point 1, $\boldsymbol{p}_{\text{fc2}}$ the position of point 2 and $\boldsymbol{p}_{\text{fc3}}$ point 3.

First, the translation $\boldsymbol{t}_{\text{fc}}$ was determined as the center of the points

$$\boldsymbol{t}_{\text{fc}} = \begin{bmatrix} \text{mean}(x_{\text{fc1}}, x_{\text{fc2}}, x_{\text{fc3}}) \\ \text{mean}(y_{\text{fc1}}, y_{\text{fc2}}, y_{\text{fc3}}) \\ \text{mean}(z_{\text{fc1}}, z_{\text{fc2}}, z_{\text{fc3}}) \end{bmatrix}. \tag{3.2}$$

Then a new coordinate system illustrated in Fig. 3.5 was created. The three points $\boldsymbol{p}_{\text{fc1}}$, $\boldsymbol{p}_{\text{fc2}}$, and $\boldsymbol{p}_{\text{fc3}}$ create a plane. The normal of this plane was chosen as the z-axis, computed as

$$\boldsymbol{z}_{\text{fc}} = (\boldsymbol{p}_{\text{fc2}} - \boldsymbol{p}_{\text{fc1}}) \times (\boldsymbol{p}_{\text{fc3}} - \boldsymbol{p}_{\text{fc1}}). \tag{3.3}$$

The x-axis corresponds to a vector begining in the center and crossing through the point $\boldsymbol{p}_{\text{fc1}}$

$$\boldsymbol{x}_{\text{fc}} = \boldsymbol{t}_{\text{fc}} - \boldsymbol{p}_{\text{fc1}}. \tag{3.4}$$

The y-axis in a right-handed coordinate system can be computed by rotating the x-axis by the z-axis by 90°. This was achieved by creating a transformation matrix in Matlab using makehgtform('axisrotate', $\boldsymbol{z}_{\text{fc}}$, $\frac{\pi}{2}$) and then transforming the x-axis $\boldsymbol{x}_{\text{fc}}$.

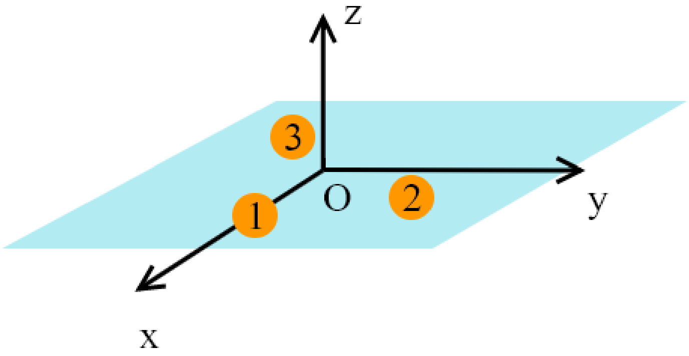


**Figure 3.5:** A new coordinate system of a rigid body described by points 1, 2 and 3 in a plane (blue). The origin is the middle of the points, z-axis is the normal of the plane defined by the three points, the x-axis goes through point 1 and the y-axis is the x-axis rotated by the z-axis by 90°.

To generate a rotation matrix, the axis vectors must be normalized and then ordered into a 3x3 matrix as follows

$$\boldsymbol{R}_{\mathrm{fc}} = \begin{bmatrix} \frac{\boldsymbol{x}_{\mathrm{fc}}}{\|\boldsymbol{x}_{\mathrm{fc}}\|} & \frac{\boldsymbol{y}_{\mathrm{fc}}}{\|\boldsymbol{y}_{\mathrm{fc}}\|} & \frac{\boldsymbol{z}_{\mathrm{fc}}}{\|\boldsymbol{z}_{\mathrm{fc}}\|} \end{bmatrix} . \tag{3.5}$$

The rotation matrices for hands and feet are not used in the joint center estimation method discussed in Sec. 3.2.1. Only the position is needed. It is calculated the same way as for all other clusters using Eq. (3.2), without the need to have three markers.

### Joint center estimation

Estimating the joint centers is accomplished by fixing a base body segment in time and space. Applying the rotation and translation calculated in Sec. 3.2.1, the position of a moving body segment is transformed into the coordinate system of the base body part. For example, to estimate the wrist position, the base segment would be the forearm and the moving segment would be the hand.

The following formula was used to transform the moving positions $\boldsymbol{t}_{\mathrm{fc,moving}}$ from the general coordinate system O into the fixed coordinate system B of the base body part

$$\boldsymbol{t}^{\mathrm{B}}_{\mathrm{fc,moving}} = (\boldsymbol{t}^{\mathrm{O}}_{\mathrm{fc,moving}} - \boldsymbol{t}^{\mathrm{O}}_{\mathrm{fc,base}}) \cdot \boldsymbol{R}^{\mathrm{O}}_{\mathrm{fc,base}} . \tag{3.6}$$

These transformed points $\boldsymbol{t}^{\mathrm{B}}_{\mathrm{fc,moving}}$ create spheres around the joint center. The center and radius of this sphere is obtained by a Matlab implementation [41] of the Pratt sphere fitting method [42]. The process is illustrated in Fig. 3.6.

The joint centers $\boldsymbol{c}^{\mathrm{B}}_{\mathrm{fc}}$ are then transformed back into the general coordinate system using

$$\boldsymbol{c}^{\mathrm{O}}_{\mathrm{fc}} = \boldsymbol{c}^{\mathrm{B}}_{\mathrm{fc}} \cdot (\boldsymbol{R}^{\mathrm{O}}_{\mathrm{fc,base}})^{-1} + \boldsymbol{t}^{\mathrm{O}}_{\mathrm{fc,base}} . \tag{3.7}$$

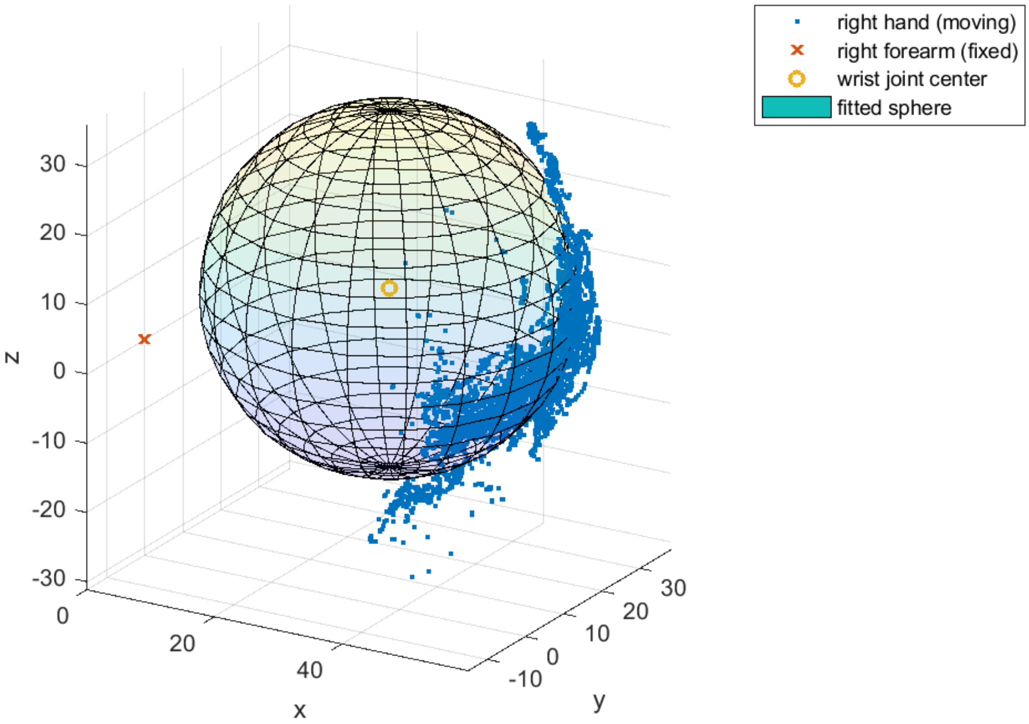


**Figure 3.6:** Example of joint center estimation. The moving points follow the shape of the sphere, that is fitted to them. The joint center is the center of the sphere.

### 3.2.2 Data synchronization

Two systems to collect data were used: the Qualisys motion capture system and two RGB-D cameras. The data from these sources are in different coordinate systems and framerates. Also, the recording starts at different times, meaning they have a different time axis. Qualisys RGB and motion capture is synchronized internally, but the RealSense cameras are not. The synchronization process is described in the following sections.

#### Timeframe synchronization

To reliably synchronize the footage, a frame that corresponds to the same time must be found in all data sources. Since the Qualisys video camera is synchronized with the motion capture system, only the Qualisys RGB and RGB-D footage must be synchronized. Zhao et al. [33] rely on a starting pose and match the frames using a cost function for synchronization. Infants cannot be instructed to strike a pose for synchronization purposes. A flash of light was used after recording started. The first frame of the footage, where the flash was no longer visible, was matched through the data recordings.

### Unification of frame rates

The motion capture data was mostly recorded at 100 fps, while the RGB video from the same system at 25 fps. To unify the framerates, every fourth sample of the motion capture recording was used, while the rest was left unused.

The RGB-D camera records with a framerate $f_{\mathrm{rgbd}}$ varying through the experiments. To find the corresponding RGB-D frame for the Qualisys video with 25 fps, the number of the Qualisys frame was multiplied by $f_{\mathrm{rgbd}}/25$ and then rounded to the nearest integer.

### 3.2.3 Combining the pointclouds from RGB-D cameras and motion capture keypoints

To utilize the information from all the data sources, alignment had to be computed. This section describes the process of obtaining transformation matrixes to transform the data to one coordinate system. The transformation of the two RGB-D cameras is depicted in Fig. 3.7 while the

A calibration board with markers secured to it was used for the alignment. The alignment was done by comparing the board plane equations in format

$$\boldsymbol{r} \cdot \boldsymbol{n} = \boldsymbol{d}\,, \tag{3.8}$$

where $\boldsymbol{n}$ is the normal of the plane and $\boldsymbol{d}$ is the distance of the plane from the origin and $\boldsymbol{r}$ is a point belonging to the plane. A picture of the board must be taken in at least three different positions to create a transformation matrix between the three camera systems (two RGB-D and optical motion capture). For better accuracy, seven board position were recorded in our trials.

As the calibration board is mainly white and no other parts of the image are, white sections of the RGB part of the pointcloud were extracted in Matlab and planes were fitted to the obtained pointclouds with a maximum distance relative to the camera distance. These plans contain only the points of the calibration boards. The Matlab command `pcfitplane()` returns the plane parameters, which were used in the later alignment.

In motion capture, four markers were detected. From these four markers the plane Eq. (3.8) is extracted. First, a normal $\boldsymbol{n}$ is computed as the normed cross product of two vectors in the planes

$$\boldsymbol{n} = \left\| (\boldsymbol{p}_2 - \boldsymbol{p}_3) \times (\boldsymbol{p}_3 - \boldsymbol{p}_1) \right\| \,, \tag{3.9}$$

where $\boldsymbol{p}_i$ are the individual markers in the corners of the calibration board. The parameter $\boldsymbol{d}$ is computed substituting $\boldsymbol{n}$ and a point $\boldsymbol{p}_1$ in Eq. (3.8) as follows

$$\boldsymbol{d} = \boldsymbol{r} \cdot \boldsymbol{n} = \boldsymbol{p} \cdot \boldsymbol{n}\,. \tag{3.10}$$

After all the plane equation parameters are calculated, the planes are rotated in the same direction and the rotation matrixes between the planes are computed. Then the data is scaled by 1000, as RGB-D data is in metres, while the motion capture is in milimeters. By comparing the median as the center of all points a translation is computed.

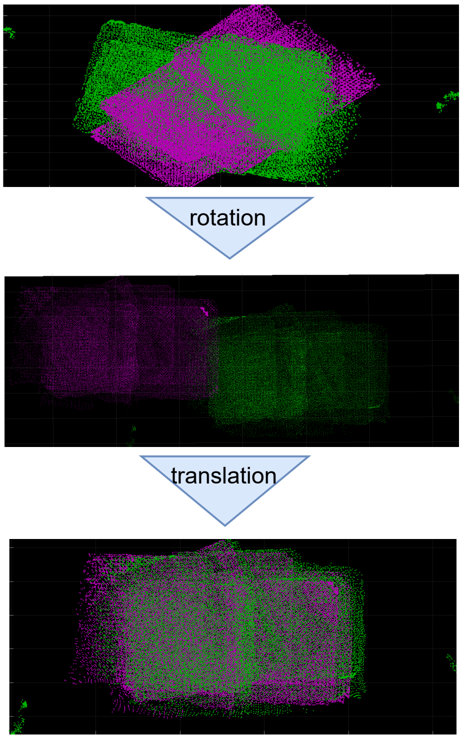


**Figure 3.7:** The board captured by the RGB-D camera is extracted (above), then the two planes are rotated to face the same direction (middle) and last, the boards are translated so that the middle points are in the same position.

### 3.2.4 Combining the pointclouds from RGB-D cameras and SMIL keypoints

SMIL provides 3D meshes of the infants, while the RGB-D cameras output pointclouds. These two can be combined and the information compared. It was accomplished as follows.

The background behind the infant in the RGB-D footage is flat, therefore it coud be removed by using plane extraction. The 3D meshes were converted to pointclouds that were manually overlayed on the RGB-D infant pointclouds using rotation and translation. This positioning was the initial position before ICP (iterative closest point) was run [43].

## 3.3 Evaluation of joint extraction

Metrics are used to evaluate the precision of estimating joints. Most evaluation metrics use the comparison of estimated keypoints against ground truth data. However, to evaluate the accuracy of the ground truth motion capture data, we have no other ground truth to compare it to. Therefore, standard bone deviation was chosen to signify the precision of our method, as it is monitoring the changes in lengths of each bone. The next chosen method is the mean per joint position error, which compares the absolute positions of joint centers. To evaluate the data using this method, it must be in the same coordinate system and scale. The unification of the keypoints and evaluation metrics are described in the following sections.

### 3.3.1 Unifying the skeleton joints and their positions

The skeleton keypoints from our method varies from the output from SMPLify-X with SMIL, and possibly other skeletons. SMPLify-X has 25 keypoints, whereas our skeleton model has only 14. For comparison purposes, the extra keypoints like eyes, ears and toes were not accounted for.

Then the scale and origin of the skeleton coordinate systems of the two skeletons had to be unified. The data from SMPLify-X with SMIL went through the same normalization process as in [5], with the middle point between the hips fixed at (0,0,-1) and the middle point between shoulders at (0,0,0). The rotation of the body was determined by the vector from hip to hip. To accomplish the transformation to one unified coordinate system, transformation matrixes were created and applied to bothe the motion capture data and the model data.

### 3.3.2 Bone standard deviation

Bones in the skeleton models are structures connecting joints as in the human body.Bones should not deviate in length, therefore, to show the accuracy of a joint estimation model, the standard deviation of the bone lengths is computed.

The length $d$ of a bone defined by two joints with positions

$$j_1 = \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}, j_2 = \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} \tag{3.11}$$

and is computed as

$$d = \|j_1 - j_2\| . \tag{3.12}$$

Then the standard deviation is calculated using Matlabs std() command. To compare the bone standard deviation between the two skeletons, they have to be unified as explained in Sec. 3.3.1.

### 3.3.3 Mean per joint position error

Mean per joint position error compares the absolute position of joint centers of ground truth and estimated joints. Therefore, the data must be canonized as described in Sec. 3.3.1. For each frame $f$ a value $E_{\mathrm{MPJPE}}$ is calculated using an euclidean norm as follows

$$E_{\mathrm{MPJPE}}(f) = \frac{1}{N} \sum_{i=1}^{N} \|P_{\mathrm{e}}(i, f) - P_{\mathrm{gt}}(i, f)\| , \tag{3.13}$$

where $N$ is the number of joints, $P_{\mathrm{e}}(i, f)$ is position of estimated joint $i$ in frame $f$ and $P_{\mathrm{gt}}(i, f)$ is position of ground truth joint $i$ in frame $f$.

## 3.4 Safety of infants during recordings

The experiments were approved by the Committee for Research Ethics at the Czech Technical University in Prague (ref. no 0000-02/23/51902/EKCVUT) followed guidelines granted by the Committee for Ethics in Research of the Czech Technical University in Prague. The parents signed an informed consent form. Some precautions are discussed in the following sections.

### 3.4.1 Cameras

The RGB-D cameras emit low intensity infrared light that poses no risk to infants. In our experiments with infants, we use the Intel RealSense D455 camera. Its predecessor, the Intel RealSense SR300 camera has been used in intensive care units [18] even with preterm neonates [44]. The motion capture system is certified for use on humans.

### 3.4.2 Risks of swallowing markers

The motion capture markers were 9.5 and 19 mm in diameter. The size poses a risk of swallowing. This risk was mitigated in two ways. First, the markers were never put on the infant individually. They were in groups of two to three, as described in Sec. 4.1.7. Second, at least one researcher was was focused directly on the markers. If the infant took any off, the researcher would immediately remove the marker from the infant. In addition, the utilization of marker clusters (see Sec. 4.1.7) served to minimize risks. The markers were securely affixed in groups of two or three using a combination of tape and fabric, resulting in an increased adhesive surface area.

# Chapter 4

# Experiments and Results

## 4.1 Preparation for the infant recording

To ensure maximum safety and comprehensive knowledge of the setup before recording infants, an adult, a stylized baby doll and a realistic animated baby doll were recorded in several sessions. The preparatory experiments are described in the following sections. Their goals and findings are summarized in Table 4.1.

| Subject | Goal | What was learned | Section |
|---|---|---|---|
| adult | test the Qualisys Animation markerset skeleton | skeleton extraction is functional, not appropreate for infants | 4.1.1 |
| stylized doll | try the Qualiys sports animation marker set | same problems as Animation marker set | 4.1.2 |
| stylized doll | try spherical markers and retroreflexive tape | spherical markers work better | 4.1.2 |
| stylized doll | try elastic bands for marker adhesion | the bands are not suitable for infants | 4.1.2 |
| stylized doll | try to combine motion capture and rgbd | the two cameras do not interfere with each other | 4.1.3 |
| realistic doll | test marker size and stands | the optimal solution is a marker with diameter 0.9 cm on a rigid stand | 4.1.4 |
| realistic doll | try sphere fitting joint estimation | it is suitable for infants | 4.1.5 |
| adult | test sphere fitting against anatomical landmarks | sphere fitting is more reliable | 4.1.6 |
| adult | try setup and introduce clusters | the setup is ok, clusters are great | 4.1.7 |

**Table 4.1:** Summary of preparatory experiments.

### 4.1.1 Adult Animation marker set

First, to try the motion capture system, an adult was recorded mimicking the motion of an infant while standing. This section describes the setup and results of the experiment.

### Marker setup

To capture the motion, the Qualisys Animation marker set was used. It is based on 44 markers secured on anatomical landmarks of the body. The software for capturing and processing the motion capture data, the Qualisys Track Manager (QTM), provides a functionality that generates a skeleton model from this particular set of markers. The skeleton is shown in Fig. 4.1a.

Double-sided skin-safe tape was used to place all 44 markers. The process to palpate the correct landmark using the Anatomical Atlas [25] and adhere the markers to the skin took around 30 minutes.

### Motion capture setup

Eight motion capture cameras were used around the perimeter of our laboratory, as described in Sec. 3.1.1. The Qualisys RGB camera was mounted on a tripod to capture the front view of the moving adult.

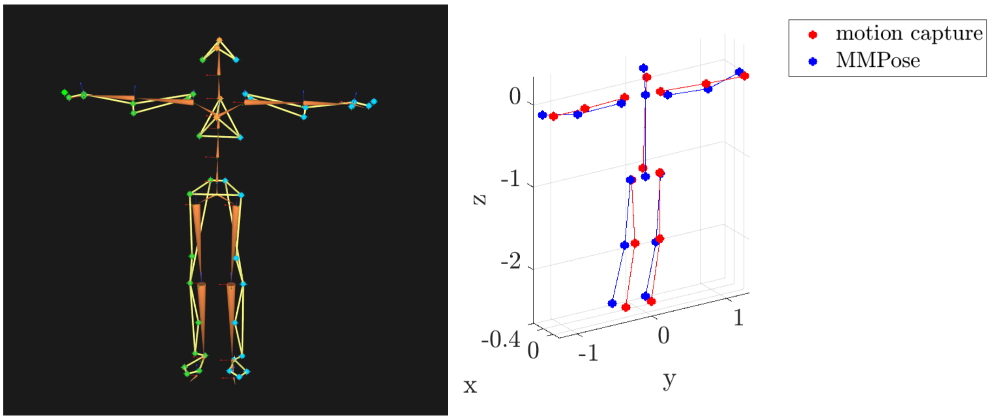
The auto-exposure and auto-white balance of the video camera was turned on. The marker threshold was set at 40 %, while the exposure and flash time was at 300 $\mu$s. The sampling rate of the motion capture cameras was at 100 fps.

A problem while setting up the motion capture arised: Markers were detected by the motion capture system that were not present in the scene. These phantom markers were created by reflections of the light from reflective materials other than the markers. There are two solutions to solving this problem: The phantom markers can be masked in QTM or located in the room and removed or covered. Masking is selecting a group of pixels from the infrared camera that are not used in the triangulation of markers. QTM provides an auto-masking feature that masks every reflected spot in the view. As glass is a highly reflective material, the windows had to be closed and the blind partially closed for all the experiments.

### Comparison with MMPose keypoints

The skeleton data from QTM was exported to Matlab and compared with an MMPose [45] estimation of movement. The MMPose model has 25 keypoints, whereas the Qualisys Animation has 18. Keypoints like nose, ears and toes were left out from the plots and comparisons, as there is no motion capture data to compare them with. The remaining keypoints were unified in scale using the method explained in Sec. 3.3.1. Differences in the joint positions of the two models are illustrated in Fig. 4.1b.

Skeletal models were evaluated using *MPJPE* and bone standard deviation (described in Sec. 3.3.3). Figure 4.2a depicts the *MPJPE*. The spike around frame 3000 could be explained by a complex movement of the adult's body. Otherwise, the joint position is relatively similar. An interesting thing to point out is the beginning and end of the recording, where the person struck a synchronization T pose. The error does not deviate much in these frames. The bone standard deviation is calculated after fixing the neck and hips, to allow for a scaled comparison, as shown in Fig. 4.2b.



**(a)** The skeleton from the Qualisys animation marker set is represented by the orange cones, the markers are blue (right side) and green (left side).

**(b)** The comparison of the same keypoints from MMPose and the Animation marker set. The keypoints are canonized the neck and midhip as explained in Sec. 3.3.1.

**Figure 4.1:** The skeletons obtained using the Animation markerset compared with MMPose.

### Results

There are three main disadvantages to this method of capturing movement. The animation marker set requires seven markers on the back and two on the heels, which is not possible to capture on an infant in supine position. Second, the marker placement takes around 30 minutes, which is too long for an infant. Young infants can be fussy and easily tired, therefore, the process of marker setup has to be very fast. Lastly, the Animation skeleton is, as the name suggests, intended for animation, not motion analysis.

The results of the skeleton comparison with MMPose are not surprising, as both methods were developed for adult motion tracking.

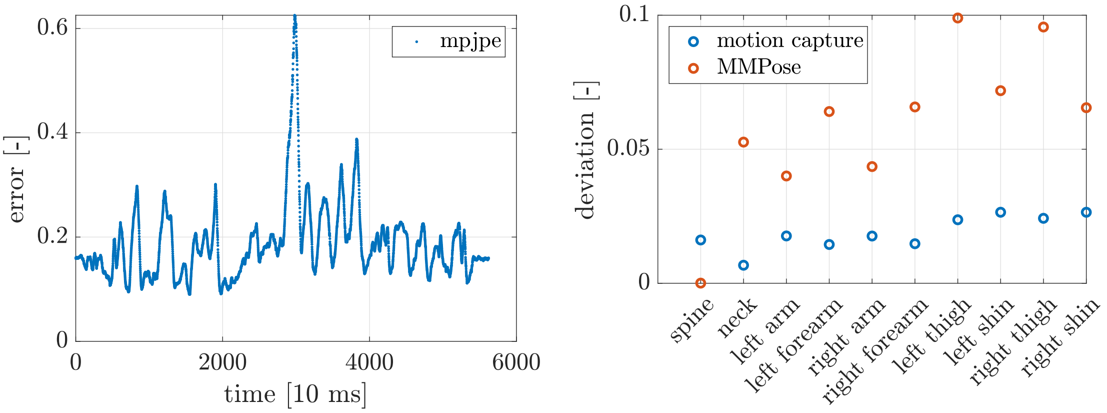
In this experiment, the Animation marker set was tested and found to be unsuitable for usage on lying infants. The evaluation methods were used without problems.

## 4.1.2 Stylized doll sports marker set and marker type

The objective of this experiment was to test the skeleton of the Sports marker set provided by Qualisys and to try different marker types and adhesion solutions. A doll (in Fig. 4.3a), approximately 60 cm tall, was used to represent the size of the infant. The doll did not have articulated joints, so there was a limited possibility of mimicking infant movement.

### Marker setup

The Sports marker set should be more equipped for motion analysis than the Animation marker set mentioned in Sec. 4.1.1. The skeleton model and marker layout can be seen in Fig. 4.3a. Similarly to the Animation marker set, the Sports marker set requires markers on the back side of the body, which is not suitable for an infant in a supine position.



**(a)** *MPJPE* of the MMPose skeleton against the Animation motion capture data set as ground truth in time.

**(b)** Comparison of the mean of standard bone deviation of the motion capture ground truth and MMPose.

**Figure 4.2:** The evaluation of motion capture joint extraction and the MMPose extraction from a video. The deviation is computed on the canonized data (see Sec. 3.3.1), therefore, the error an deviation do not have a unit.

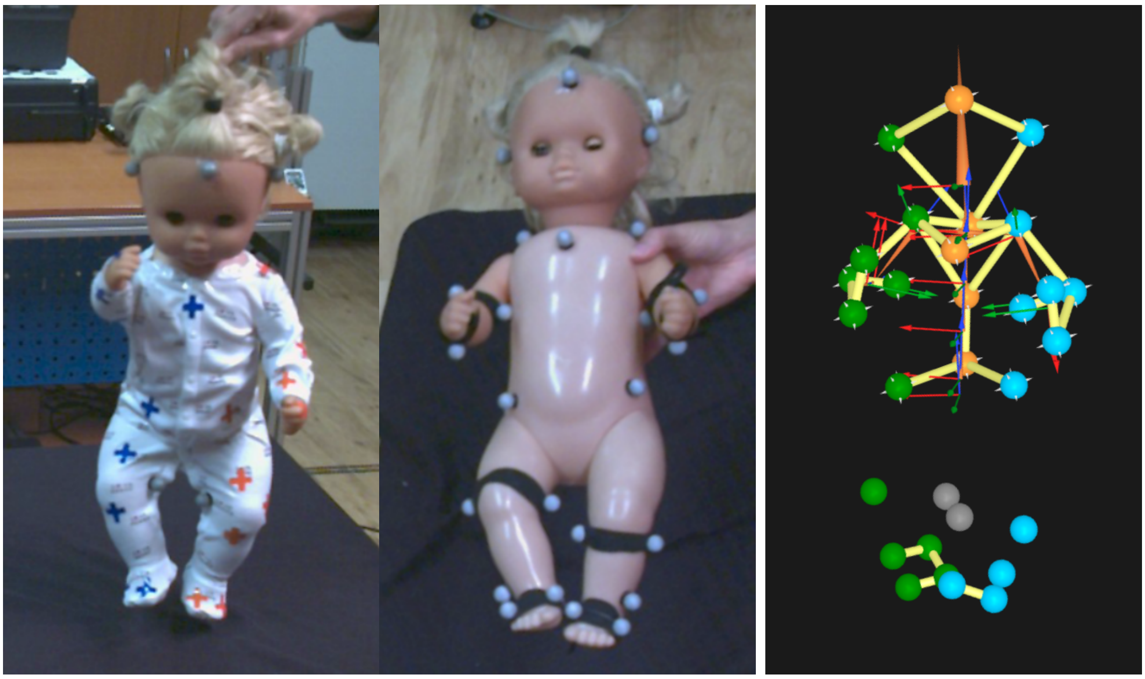
Two types of markers were available and tested: spherical markers with 1.9 cm diameter and retroreflective tape. The tape had been cut into crosses, because it is an easy shape to manufacture with limited equipment and it wraps nicely around the limbs of a small body. To find which marker type worked better in tracking of infant movement, they were put on the doll referencing the Sports marker set guide and the doll was moved in space. The recognition of markers by the motion capture system was visually observed. Confirmation of the information from Sec. 3.1.1, the spherical markers were deemed more reliable than the retroreflective tape.

In addition, a new marker attachment method was developed as an alternative to double-sided tape. The spherical markers were placed in holes in an elastic with buttonholes. The elastic was then wrapped around a limb and secured with a hook, sewn onto one end. The band with markers is shown in Fig. 4.4.

The elastic band solution (Fig. 4.4) for marker attachment has two main flaws. First, the band would be directly on a moving joint, possibly restricting the infant's movement. Second, the elastic can be tightened in increments of 1.5 cm, meaning the fit could be too loose or too tight on the infants limbs. If the fit were too loose, the markers could shift in position, ruining the recording. If the fit was too tight, it could be uncomfortable for the infant and possibly even cut its blood circulation.

### Motion capture setup

The motion capture cameras were in the formation described in Sec. 3.1.1. The video camera was mounted on a tripod about 1.5 meters above ground, pointed in the middle of the room, where the doll was being moved. The auto-exposure and auto-white balance were on. The marker threshodl was set at 20 % while the exposure and flash time were 300 $\mu$ s.



**(a)** The doll with crosses cut from retroreflective tape to serve as markers.

**(b)** The doll with spherical markers secured with elastic bands.

**(c)** The torso of the Qualisys Sports skeleton is represented by the orange cones.

**Figure 4.3:** Marker setup and configuration for testing the retroreflective tape and spherical markers.

### Results

These experiments have proven the unsuitability of the Sports marker set for infant recording due to the inability to capture markers on a lying infant's back. Spherical markers were picked against the cut retroreflective tape due to better visibility for more cameras. A new marker attachment solution was proposed, but rejected for further use as it could restict the infants movement and blodflow to the limbs.

## 4.1.3 Combining motion capture and RGB-D

In this experiment, the combination of optical motion capture and a RGB-D camera for data acquisition was tested. The main objective was to make sure that the two camera systems do not obstruct each other, as both camera systems use infa-red light.

### Marker setup

The marker setup consisted of 23 markers on the body as shown in Fig. 4.5a. No joint extraction method was used on this set of markers, it was used only temporarily to test the compatibility of motion capture and RGB-D .

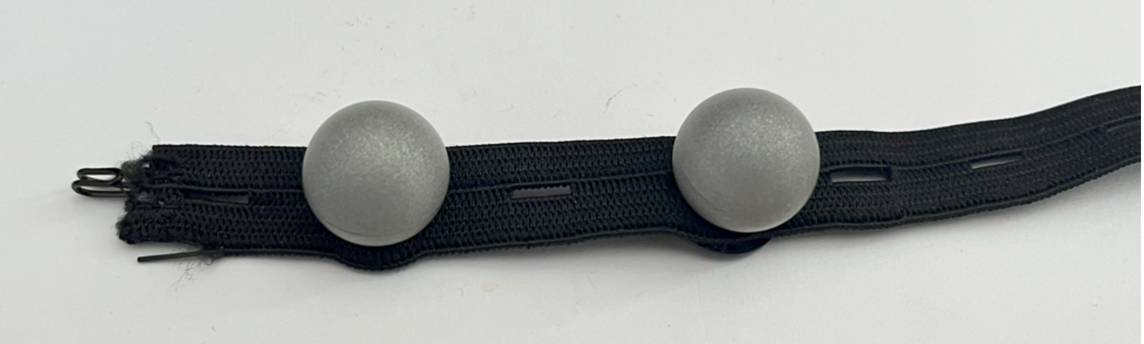


**Figure 4.4:** An elastic band with buttonholes with two markers secured and a hook for attachement.

### Motion capture setup

The RGB data was obtained from the RGB-D camera, therefore the Miqus video camera was not used. The infra-red cameras were set as described in Sec. 3.1.1 with exposure and flash time at 300 $\mu$s, marker threshold at 20 % and framerate at 50 fps.

### RGB-D camera setup

One Intel RealSense D435 camera was used in this experiment. It was mounted on a rig suspended from the ceiling approximately 1.5 meters above ground. The camera settings were set as follows: the defaul camera mode was picked, all postprocessing was disabled to obtain the data in a raw form, auto-exposure was diabled for the depth module, and enabled for the RGB information. The resolution of the depth module was 848 x 480, while RGB had resolution at 1280 x 720. Both the depth and RGB data was captured with the framerate at 30 fps.

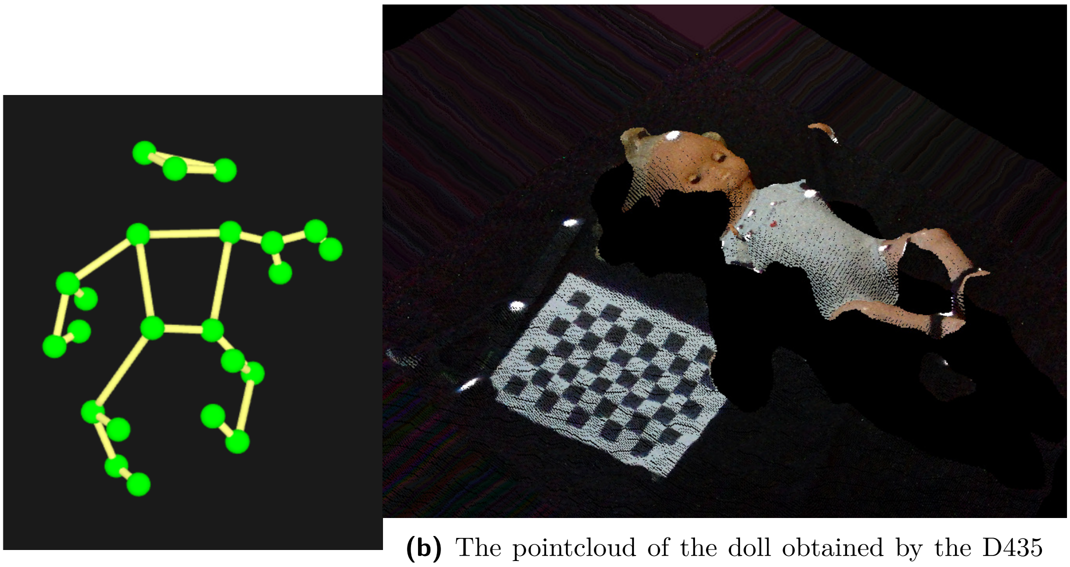
An example of the captured pointcloud is in Fig. 4.5. The shadows in the point cloud is missing information, that could not be recorded because the camera cannot see through objects. To minimize these shadows, more RGB-cameras can be used.

### Results

This experiment showed that there are no problems in recording with the motion capture system and the RGB-D cameras at the same time. No interference between the two camera systems or unwanted reflections were discovered. The RGB-D pointclouds had many holes, because the camera cannot see behind the limbs; therefore, it would be better to use multiple RGB-D cameras. For synchronization purposes, it would be better to use the motion capture system with the synchronized RGB video camera, as it is not possible to use the flash to synchronize the motion capture data with RGB-D without it.

### 4.1.4 Realistic doll marker size test

In previous experiments (see Sec. 4.1.2), spherical markers were determined to be a preferable alternative to retro-reflective tape. The objective of this particular experiment



**(a)** The marker setup.

**(b)** The pointcloud of the doll obtained by the D435 camera.

**Figure 4.5:** The recorded motion capture markers on the doll were not used for joint estimation, only to test motion capture and RGB-D camera compatibility.

was to select the optimal size for the spherical markers and establish a minimal distance between two markers. Only the motion capture camera system was used for this experiment, as the RGB-D footage is not affected by the size of the markers.

### Marker setup

Markers with diameters of 0.6 cm, 0.95 cm and 1.9 cm (in Fig. 4.6) were adhered to the doll,which was moved and recorded. Visual evaluation indicated that the behaviors of medium and large markers were quite similar, while the smallest markers were occasionally not recognized.

Considering the size of infants, smaller markers would be the most suitable option. However, their reliability was not consistent, leading to the use of medium-sized markers. As discussed with Hoshinori Kanazawa, this probably has to do with how far the motion capture cameras are from the recorded areas in our setup (see Sec. 3.1.1); if they were closer as in his own setup [31], then smaller markers would probably be recognized. Unfortunately, it is currently impossible for us to set up our cameras so close to the infant.

The medium markers were equipped with rubber stands, making it possible for infants to remove them, which poses a risk of ingestion. To address this concern, the rigid stands designed for the large markers were combined with the medium markers. To investigate further, a new experiment was conducted to evaluate potential differences between the performance of medium markers on the taller rubber stands, medium markers on the shorter rigid stands and small markers (see Fig. 4.6).

Three markers of each type (small, medium on rubber stands, medium on rigid stands, see Fig. 4.6) were placed on a cardboard plane at distances of 3 cm, 2 cm, and then 1.5 cm.

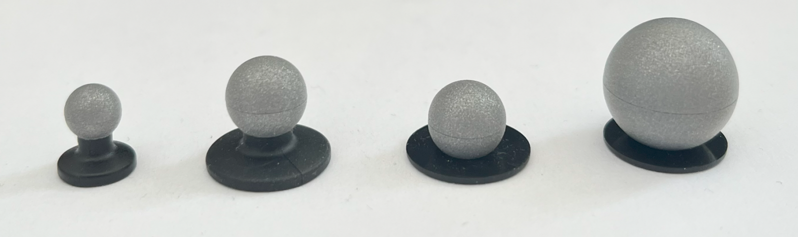


**Figure 4.6:** Markers with diameters of 0.6 cm, 0.95 cm on the rubber stand provided by the supplier, 0.95 cm on a rigid stand and 1.9 cm.

The plane was slowly turned upside down three times while being recorded by the motion capture system.

The data obtained was visually evaluated, focusing on two main objectives: the angle of rotation of the board at which markers were no longer recognized by the motion capture system and how the markers shifted positions relative to each other. Variations observed between marker types were minimal.

#### Motion capture settings

The Miqus video camera was mounted on a tripod, directed above the volume of interrest. Auto white balance and auto exposure were on. The threshold was 20 %, exposure time 300 $\mu$s and framerate 50 fps.

#### Results

Based on the findings of this experiment, medium markers with rigid stands were selected for infant recordings, as they cannot be easily detached from the stands. Distances of 2 cm and 3 cm yielded satisfactory results, while 1.5 cm was too close, causing marker merging. Therefore, the minimal distance between two markers was established at 2 cm.

### 4.1.5 Spherical joint estimation for realistic doll

This experiment was conducted to verify the applicability of the joint estimation method based on spherical fitting explained in Sec. 3.2.1. The benefit of this method is that the markers do not have to be placed precisely, but on a general area. The realistic animated doll was used to mimic spontaneous infant movement. Clear strings were tied to its limbs, so it could be puppeteered without a person present and obstructing markers from the motion capture cameras and the RGB videos.

#### Marker setup

To detect individual limbs, markers were placed individually on the infant's body in the formation shown in Fig. 4.7 inspired by Kanazawa et al. [31]. Each body part was defined by one cluster of three markers, which made it possible to define rigid bodies as described in Sec. 3.2.1 and then estimate the joint positions.

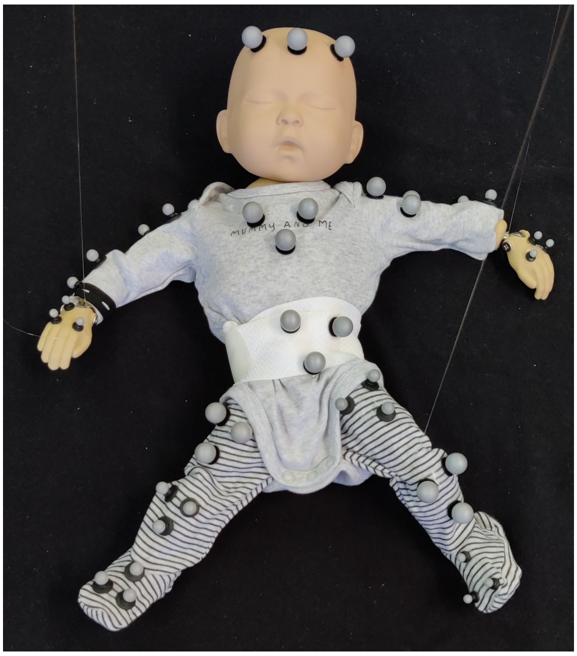


**Figure 4.7:** The layout of marker clusters secured to the doll individually using tape. The different sizes of the markers were irrelevant for this experiment.

#### Motion capture settings

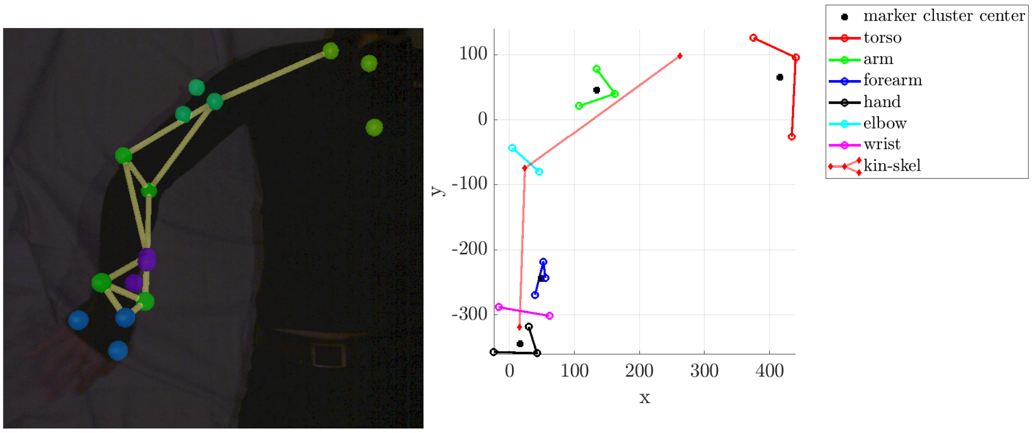
The video camera was mounted on a rig suspended from the ceiling, overlooking the volume of interrest from above. Auto-exposure was turned off while auto-white balance was left on. The exposure and flash time of the infrared cameras was at 214 $\mu$s, marker threshold at 20 % and sampling rate at 50 Hz.

#### Results

The main advantage of this method is it does not require markers on the back of the infant's body and they do not have to be placed precisely, which was the main difficulty with previous joint estimation solutions (4.1.1 and 4.1.2). However, sticking 45 individual markers on the infant's body would be very time-consuming, which could possibly lead to the infant being uncomfortable and fussing. In the next experiments, a better solution for marker placement was found.

### 4.1.6 Comparison of spherical joint estimation with anatomical landmarks for an adult arm

The objective of this experiment was to validate the sphere-fitting joint estimation method. Only an adult arm was recorded, as the method does not change from one limb to another. In addition, a joint estimation method based on anatomical landmarks was tested.



**(a)** Overlay of the motion captured data on the RGB video footage.

**(b)** The processed markers with estimated joint centers using the sphere fitting method.

**Figure 4.8:** The captured adult arm used to validate the use of sphere fitting for joint estimation.

### Marker setup

The arm was covered in 16 markers. Four were on the bony landmarks [25] of the wrist and elbow, and 12 markers were placed according to the sphere-fitting method, as shown in Fig. 4.8a. All markers were secured with double-sided tape.

### Comparison of sphere-fitting and anatomical landmarks

To compare the two methods of joint extraction, only two joints and one bone could be used, since anatomical landmark markers were placed on the wrist and elbow. The length of the forearm bone was chosen as the metric for comparison. As seen in Figure 4.9, the anatomical landmark solution gives a consistent result, whereas the sphere fitting has a few outliers. These outliers could be caused by marker obstruction, as there are more markers needed for the sphere fitting joint extraction. Excluding the outlier points, the results are similar in consistency. Therefore, a different criterium had to be chosen.

The palpation of anatomical landmarks on an moving infant is challlenging. Whereas placing clusters of three markers on individual body parts without the need for precision could be quicker and easier. Therefore, the joint estimation sphere fitting method was chosen for further experiments.

### Motion capture settings

The video camera was positioned on a ceiling-mounted rig, providing an overhead view of the volume of interest. Auto-exposure was disabled, while auto-white balance remained enabled. The infrared cameras utilized an exposure and flash time of 267 $\mu$s, with a marker threshold set at 20 %. The sampling rate to capture data was set at 50 fps.

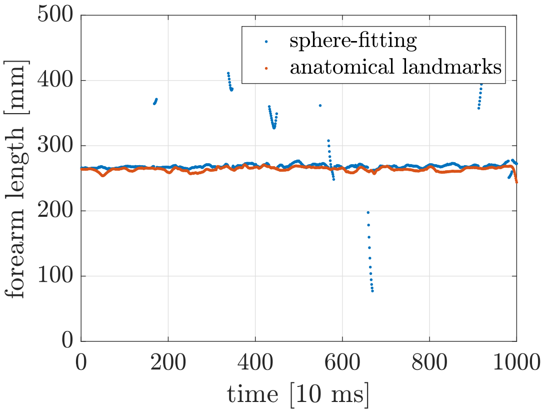


**Figure 4.9:** Comparison of forearm length estimated using joint estimation using sphere-fitting and anatomical landmarks.

### Results

The experiment has shown the superiority of the spherical joint estimation method over anatomical landmarks for our use.The only question that remains to be answered is how to stick the markers to the infant effectively.

### 4.1.7 Ádult spherical joint estimation and clusters from motion capture and RGB-D

In this final preparation experiment, the entire setup was tested and last changes were implemented. A newer model of the Intel RealSense camera, the D455, was used; markers were secured together in clusters by fabric, and some changes in their layout were made.

An adult was recorded mimicking the motion of an infant in supine position, because the movement was more natural and coordinated than the puppeted articulated doll.

### Marker setup: Clusters of markers

Clusters of two to three markers were created to be placed on the infant. The arrangement of markers in clusters ensures a constant distance between markers, which is required for creating rigid bodies in post-processing.

To speed up the process of putting markers on the subject, clusters were prepared with fabric and skin-safe double-sided tape (see Fig. 4.10). It also brings a safety feature: If the infant managed to remove them, it would not be able to swallow the individual markers.

For hands and feet, clusters with two markers connected with a strip of fabric were used. Only two markers were sufficient for joint estimation as only the position of the hands and feet is needed and not the rotation (see Sec. 3.2.1). On the head there were three markers connected by strips of fabric with medical-grade adhesive only on the stands to minimize the infant's discomfort. The rest of the clusters consisted of the same three-marker arrangement. All clusters are illustrated in Fig. 4.10.

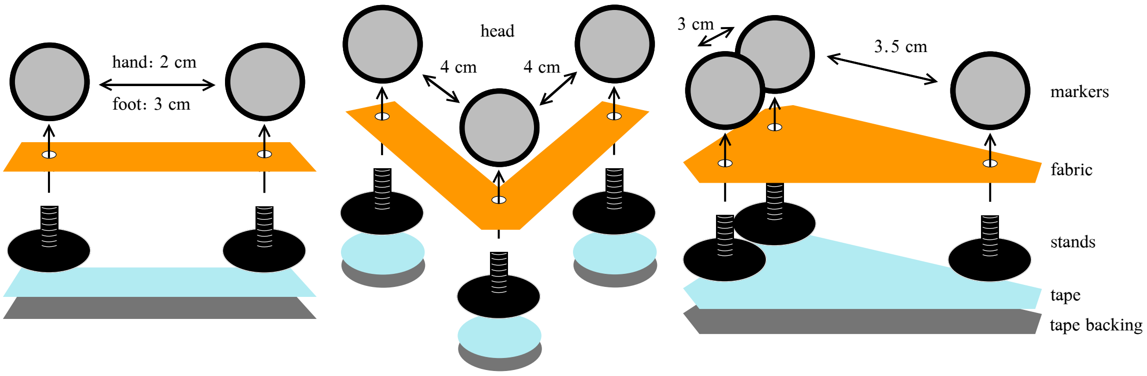


**Figure 4.10:** Different clusters held together with fabric and tape. This assembly allows for quick and easy application while preventing any choking hazard.

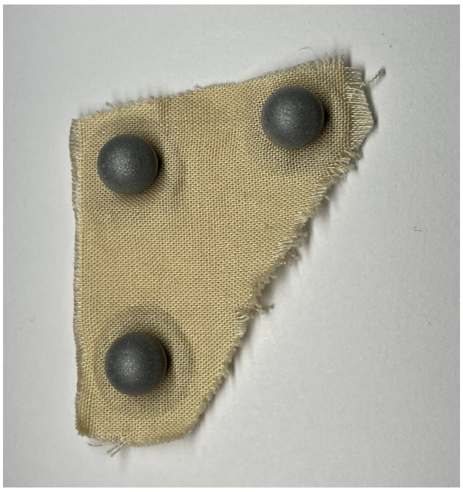


**Figure 4.11:** Cluster patch.

The layout of the markers changed slightly: on the hands and feet there were only two markers, as explained aboveand two clusters of markers were added to the back of the forearms, as an inspection of video recordings of infant spontaneous movement has shown that they tend to move their arms in an up and down motion, which would occlude the previous marker configuration. This also introduced a redundancy when defining rigid bodies that has shown to be useful in the case of occlusions.

## Motion capture settings

The video camera was positioned on a suspended rig attached to the ceiling, providing an overhead view of the volume of interest. During the recording, auto-exposure was disabled, while auto-white balance remained enabled. The infrared cameras were set with an exposure and flash time of 130 $\mu$s, a marker threshold of 15 %, and a sampling rate of 100 Hz.

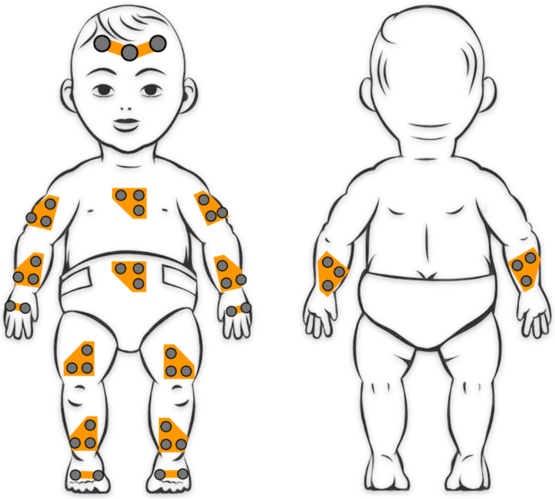


**Figure 4.12:** The infant's body was covered in 47 individual markers in 15 clusters. Clusters were strategically positioned on the limbs, aiming to maximize the distance from the base joint. For example the arm cluster was as far from the shoulder as possible.

### RGB-D camera setup

A new model of the Intel RealSense RGB-D camera, the D455, was used. The pointcloud from this camera is more detailed than the D435. Two D455 cameras were mounted on tripods approximately 1 meter above ground pointed at the volume of interest.

The parameters were set as follows: the default camera mode was picked, all postprocessing was disabled to obtain the data in a raw form, auto-exposure was diabled for the depth module, and enabled for the RGB information. The resolution of the depth module was 848 x 480, while RGB had resolution at 1280 x 720. Both the depth and RGB data was captured with a framerate of 15 fps.

### Results

This experiment confirmed that the setup and joint extraction methods were suitable for infant recordings. Cloth cluster patches were introduced, making the marker adhesion process faster and easier. The newer model of the RGB-D camera, the D455, has double the accuracy at the same distance, which allows us to place the camera a little further. The best range for D435 is between 50 cm and 1 cm; the same depth accuracy is obtained at 1m and 2m with the D455 camera [46].

## 4.2 Infant recording

Two infants, a four-month-old and an eight-month-old, have been recorded. This section describes the final setup and course of the infant recordings. All methods have been tested on a doll or an adult to make sure that the infants' time is not wasted.

### 4.2.1 Marker setup

The placement of the clusters on the infant was done from the botom to the top to go from the least uncomfortable places (foot) to the most uncomfortable places (head and hands). The mother was distracting and reassuring the infant, while two researchers placed the clusters as quickly and gently as possible to avoid infant fussiness. Due to the training of the marker placement on the doll, the entire process took approximately 3.5 minutes. The tape on the clusters was changed after each infant recording for hygienic reasons.

### 4.2.2 Procedure

This section describes how the laboratory was prepared and how the experiment was conducted. Special care had to be taken due to the live infant being the subject.

#### Laboratory preparation

First, the laboratory had to be prepared. Every surface was cleaned, floor mopped, and the robots and tables with wires that could distract or frighten the infant were covered with neutral colored blankets. Four yoga mats had to be put around a cable cover that runs through the middle of the lab, where the infant layed. Over the yoga mats, a 2 cm thick mat was placed with a washable towel on top.

The cameras were prepared as follows. The Miqus video camera was put on the rig lowered from the ceiling approximately 1.6 meters above ground. The two D455 RGB-D cameras were placed on tripods approximately 1 m above ground. The motion capture system was calibrated as described in Sec. 3.1.1. Then recordings were made with a board with markers to synchronize the two RGB-D cameras and the motion capture data (see Sec. 3.2.3). The camera settings were the same as in the last preparation experiment (see Sec. 4.1.7). The entire setup can be seen in Fig. 4.13.

#### Recording the infant

Before entering the lab, the parent was informed about the experiment procedure and risks (discussed in Sec. 3.4).

The infant was then layed on the prepared surface and markers were applied as described in Sec. 4.2.1. At this point, every person in the room went to a place where the baby could not see them to promote spontaneous movement, as that is the main research interest of the studies. Aleast one researcher was observing the infant the whole time to make sure that it did not ingest any markers. A different researcher was operating the cameras and the synchronizing flash, while a third one was watching and prepared with pieces of double-sided tape to fix the clusters, if they started to peel off.

The parent was present throughout the entire experiment. If the infant got fussy, the parent would intervene, and when the fussines persisted, the parent came to the infant's field of view and interacted with it either talking or showing them some toys.

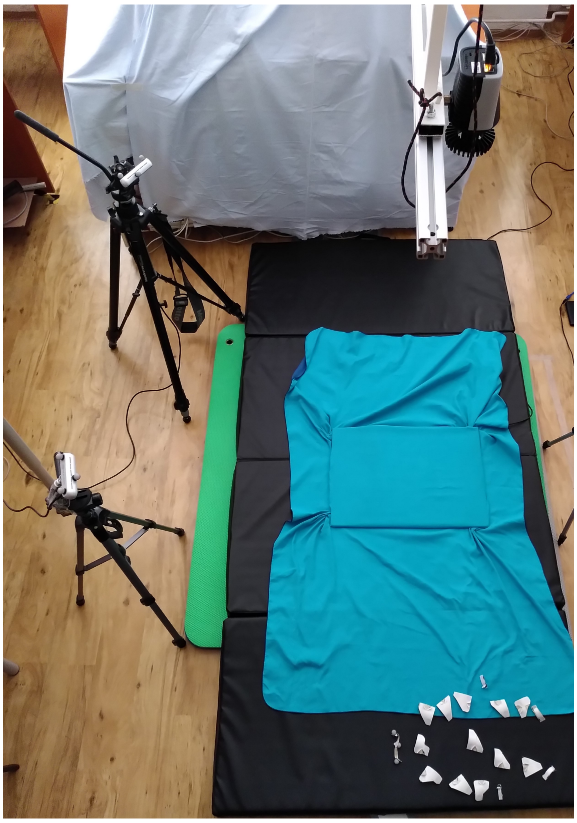


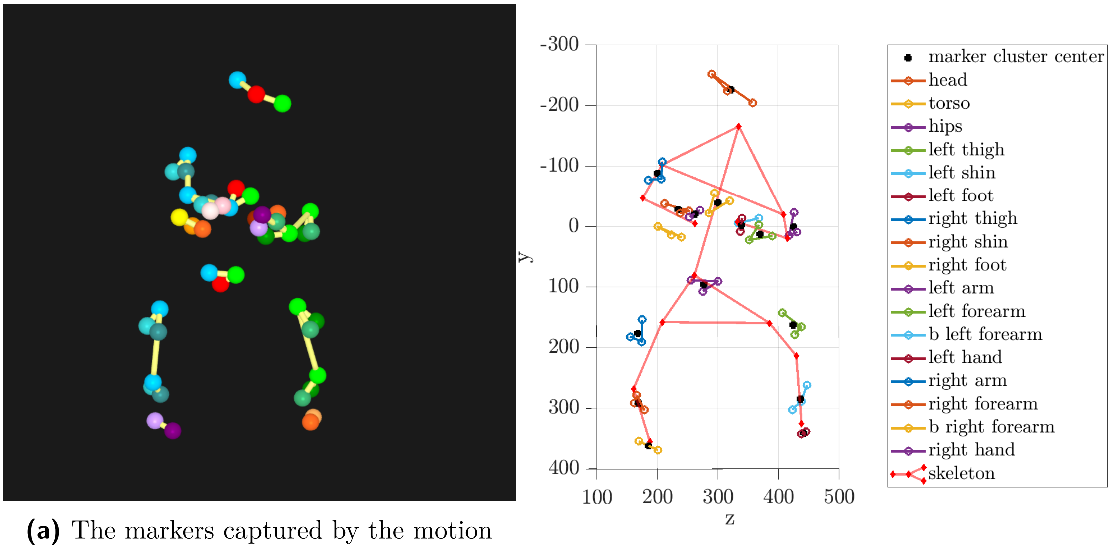
**Figure 4.13:** The setup for the infant recording consisted of two Intel RealSense D455 RGB-D cameras on tripods, an RGB Qualisys Miqus video camera mounted on a rig above the infant and eight infared cameras (not visible in this figure). The infant was positioned on a washable towel on multiple mats. In the bottom of the figure, there are prepared markers layed out in the formation they will be put on the infant.

After the recording was completed, the parent removed the marker clusters.

### 4.2.3 Data processing

The obtained motion capture data had even after labeling and gap filling gaps with missing markers. As there were not many redundant markers, an occlusion of for example a marker on the arm results in the loss of two joint positions: shoulder and elbow. To compute the joint positions (see Sec. 3.2.1), only parts of the recordings with no missing markers were used. For skeleton vizualization purposes, when there was no information on the joint position, the joint was set to (0,0,0). Otherwise, the captured data processed and synchronized as explained in Sec. 3.2 in Materials and Methods.

The joint centers calculated from the captured markers are depicted in Fig. 4.14 and in Fig. 4.15 is the combination of the motion capture and RGB-D data.



**(a)** The markers captured by the motion capture system. For reference of marker placement, see Fig. 4.12.

**(b)** The processed markers with estimated joint centers connected by bones into a skeleton.

**Figure 4.14:** Infant (eight month old) motion capture markers and joint center estimation.

### 4.2.4 Comparison with SMIL keypoints

The joint positions extracted from the motion capture in the previous section were compared with the SMIL result of the pipeline [5] (see Fig. 4.16). The data was canonized (see Sec. 3.3.1) and evaluated using the proposed metrics from Sec. 3.3. The resulting graphs are shown in Fig. 4.17. As the data in these graphs is canonized, the scale corresponds to the scale in Fig. 4.16b, therefore it could be said that the unit of the y-axis is the spine of the infant.

Both SMIL and our method performed similarly well in bone standard deviation. The sizes of individual bones fluctuate minimally in relation to the size of the spine. That could be an indication of the SMIL results being valid. However, the fact that the spine length fluctuates almost as much as the right arm, is an indication of the some wrongdoing. A comparison with the standard bone deviation with the adult motion estimation in Fig. 4.17b shows the similarity in bone deviation in the Animation skeleton and both SMIL and our method. This is a sign of positive results.

The *MPJPE* in Fig. 4.17a suggests a worse result than the adult comparison in Fig. 4.2a, as the error is overall greater for the infant estimation. This would suggest the SMIL method is not as good as the bone standard deviation would make it seem.

These results can be burdened by an error created by canonizing the data. The canonization process can propagate a small error in hip positions into a great difference of positions of the wrists. However, the canonization process is necessary, as the SMIL extraction is run on individual frames, meaning the coordinate systems and scales can and do differ. The magnitude of this problem is unknown and requires further investigation.

Overlaying the RGB-D pointcloud and mesh data (in Fig. 4.18) confirms a great problem

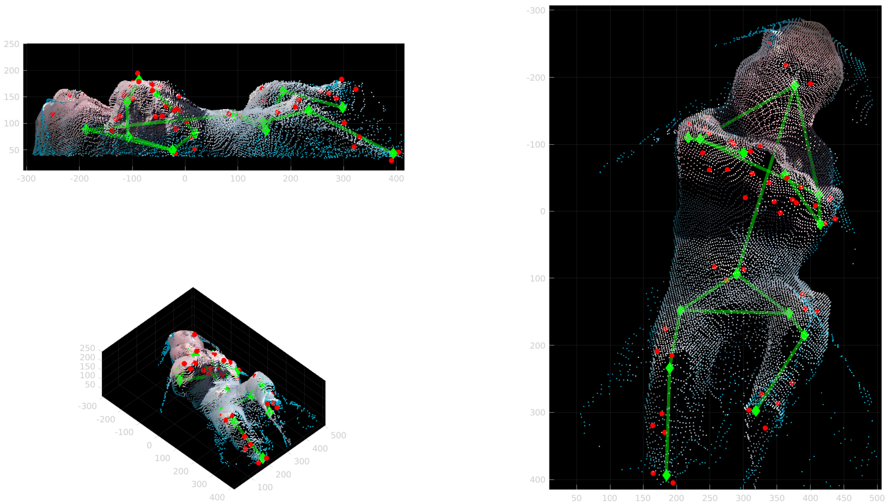


**Figure 4.15:** The skeleton and marker data of the eight mont old overlayed on the pointcloud from the two RGB-D cameras in three different views.

in pose estimation from images: The elevation of hands and feet in the direction of the camera is difficult to estimate, as the image changes only slightly with a change in elevation. The information from the RGB-D camera is used only as a visualization medium, but as many research groups show, it ha great potential. The use of RGB-D in this work is the groundwork for future prospects.

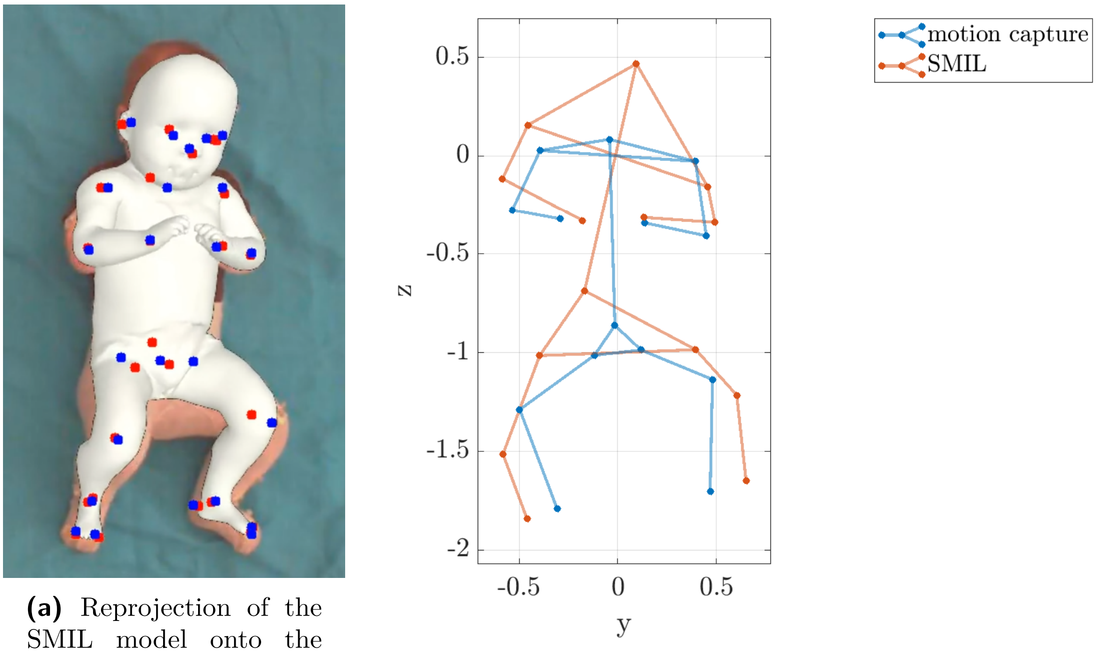
### 4.2.5 Infant recording results

The two infants were succesfully recorded using motion capture and RGB-D cameras. The joints were extracted using the sphere-fitting method and compared with the SMIL estimation. In processing of the recordings, some challenges have surfaced.

Chest markers were found to be the most frequently occluded, as both infants frequently touched their hands. Additionally, managing the large number of markers on the arms proved challenging, often resulting in marker merging or occlusions. Although increasing the number of markers could address occlusion issues, it would exacerbate the problem due to marker size relative to infant limbs. An alternative solution could involve positioning the motion capture cameras closer and using smaller markers.

Next, in the comparison of the motion capture joint centers and SMIL keypoints, it is unclear how much it influences the results of the metrics. Also, as the data is unified, no direct comparison with different results is possible.

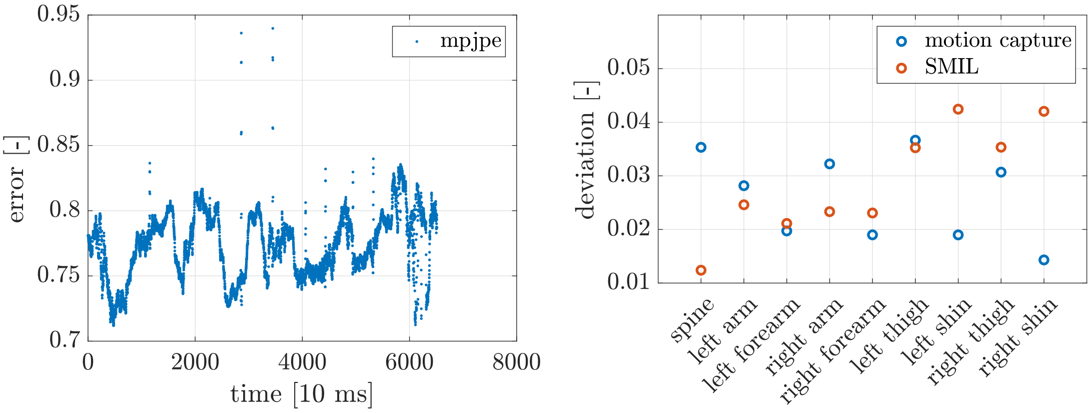
Despite the challenges, we were able to extract some infant movements and compare them with SMIL. This comparison revealed areas for improvement in all aspects of the experiment, including data acquisition, alignment, and integration of RGB-D cameras.



**(a)** Reprojection of the SMIL model onto the infant recording. [explain](NV)

**(b)** Comparison of the canonized motion capture joints and corresponding SMIL keypoints.

**Figure 4.16:** Comparison of the motion capture keypoints with SMIL of the eight month old infant.



**(a)** *MPJPE* of the SMIL skeleton against the Animation motion capture data set as ground truth in time.

**(b)** Comparison of the mean of standard bone deviation of the motion capture ground truth and SMIL.

**Figure 4.17:** The evaluation of motion capture joint extraction and the SMIL extraction from a video. The deviation is computed on the canonized data (see Sec. 3.3.1), therefore, the error an deviation do not have a propper unit.

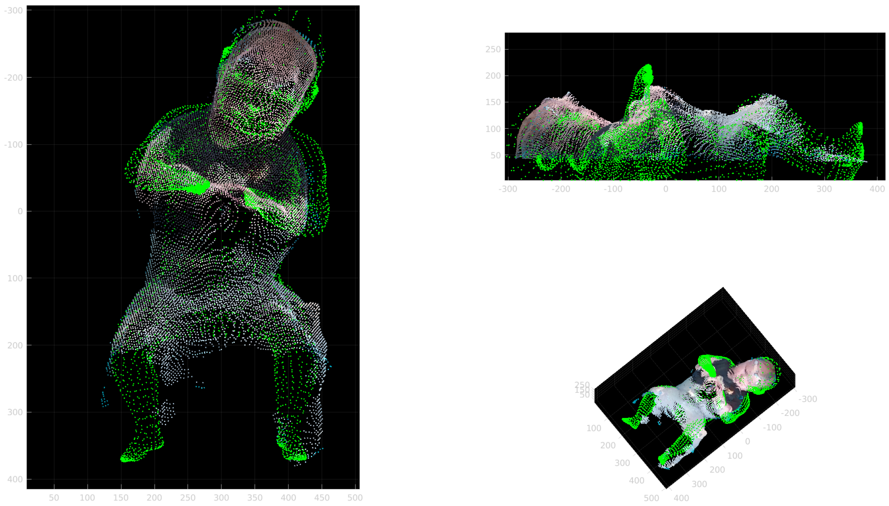


**Figure 4.18:** Comparison of the RGB-D pointcloud with the SMIL mesh (green points) of the eight month old infant from three different views.

# Chapter 5

# Discussion, Conclusion and Future Work

This work proposed a method for acquisition of ground truth of infant movement using optical motion capture and used it to make short recordings of two infants. The following paragraphs provide a summary of the achievements of this thesis as well as an outline of potential avenues for future research.

A motion capture system has been used to capture adult and infant movements. The motion capture system was created with adults in mind; therefore, all the methods and marker placement methodologies are created for adults. As infants have different body proportions and lie in the supine position in our recordings, these methods for marker placement must be modified. A marker placement method based on sphere-fitting joint estimation was proposed for our experiments.

To supplement the motion capture data, two RGB-D cameras were used to capture 3D depth point clouds with RGB information. The section of the point cloud that represents the infant was extracted from the obtained point cloud by fitting a plane to the background, as the infant is lying on the ground. Furthermore, point clouds from two RGB-D cameras were combined.

Preparation for the infant data aquisition was extensive, as seven experiments with adults and dolls were conducted to test the optimal configuration and methods to use for the infant pose estimation. However, a suit with markers has not been constructed, as the motion capture markers should not move relative to the limb they are secured to, meaning the suit should be form fitting. Furthermore, the size of infants changes rapidly; therefore, multiple different suits would have to be created, which is not cost effective. On the other hand, clusters of markers secured by fabric and tape, proposed by this work, are flexible in positioning on the body and easy to apply. There were some disadvantages to this solution, but they were only minor. The marker clusters used in the experiments were relatively large and involved significant amounts of adhesive tape, ensuring stability during recording but potentially causing discomfort during removal. Exploring the use of gentler tape to secure the marker clusters on the infant's skin while maintaining marker stability during movement could offer a more comfortable solution.

Two infants, a three-month-old and an eight-month-old, were recorded using the proposed setup. Data from these recordings showcased that there are limitation in the methods used. The primary limitation in the post-processing of the recordings was the presence of gaps in the motion capture data trajectories due to marker occlusions and marker merging. These issues stemmed from the size and placement of the markers. To address this, the implementation of smaller markers, arranged in clusters of more than three to introduce redundancy, would partially mitigate the problem.

The collected data has been temporarily synchronized using a flash of light. Then, framerates were unified at 25 fps by discarding excess frames and approximating the closest frame. This was chosen over interpolation of data, as interpolation can deform the collected data.

The obtained 3D motion capture keypoints were aligned with the point clouds from the RGB-D cameras using principles of camera calibration. Using manual transformations and ICP, the SMIL mesh was registered to the pointcloud. These two alignments help to visualize the data from multiple sources and lay the foundation for future combination and utilization of this data.

For comparison of the motion capture joint data with SMIL, two evaluation metrics were used. Mean per joint position error is a common metric that compares the absolute joint positions of two skeletons. However, to calculate the difference in joint positions, the SMIL and motion capture data must be in the same coordinate system. Therefore was the data canonized by fixing the spines in one position. This is by no means an ideal solution, as a badly identified spine could result in an error that does not correspond with reality. The second metric used is standard bone deviation. It provides the information about how much are the estimated bones in the body deviating in size.

This thesis serves as a potential starting point for further advancements in infant movement ground truth acquisition using motion capture, with the aim of creating a publicly available dataset. To achieve this, anonymization of the data would be necessary while preserving the realistic likeness of infants to ensure the authenticity of the video data set.

Improvements in motion extraction from motion capture can be pursued, potentially using human body simulators such as OpenSim [32]. This could improve joint extraction by employing accurate anatomical models of the infant's body.

Additionally, data evaluation can be improved by exploring alignment techniques, including using RGB-D cameras and mesh data for improved accuracy.

In general, these suggestions offer avenues for future research and development in infant movement analysis using motion capture technology.

# Bibliography

[1] P. Rochat, "Self-perception and action in infancy," *Experimental brain research*, vol. 123, pp. 102–109, 1998.

[2] A. Van der Meer, F. van der Weel, and D. Lee, "The functional significance of arm movements in neonates," *Science (New York, N.Y.)*, vol. 267, pp. 693–5, 03 1995.

[3] M. Barbu-Roth, "Early intervention based on neonatal crawling in very premature infants without major brain damage," 2022, accessed: 2023-05-25. [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT05278286

[4] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 12–34, 2009.

[5] J. Khoury, S. T. Popescu, F. Gama, V. Marcel, and M. Hoffmann, "Self-touch and other spontaneous behavior patterns in early infancy," in *2022 IEEE International Conference on Development and Learning (ICDL)*, 2022, pp. 148–155.

[6] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," *CoRR*, vol. abs/1611.07828, 2016.

[7] D. C. Luvizon, H. Tabia, and D. Picard, "Human pose regression by combining indirect part detection and contextual information," *Computers & Graphics*, vol. 85, pp. 15–22, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0097849319301475

[8] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei, "Human mesh recovery from monocular images via a skeleton-disentangled representation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10 2019, pp. 5348–5357.

[9] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[10] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 186–201.

[11] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, “Semantic graph convolutional networks for 3d human pose regression,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3425–3435.

[12] C. Wang, C. Kong, and S. Lucey, “Distill knowledge from nrsfm for weakly supervised 3d pose learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 743–752.

[13] B. Wandt and B. Rosenhahn, “Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7782–7791.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[15] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele, “Building statistical shape spaces for 3d human modeling,” *Pattern Recognition*, vol. 67, pp. 276–286, 2017.

[16] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, “Expressive body capture: 3d hands, face, and body from a single image,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] A. Cenci, D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, “Non-contact monitoring of preterm infants using rgb-d camera,” in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 2015, vol 9.* ASME, Design Engn Div; ASME, Comp & Informat Engn Div, 2016, aSME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Boston, MA, AUG 02-05, 2015.

[18] Y. S. Dosso, R. Selzler, K. Greenwood, J. Harrold, and J. R. Green, “Rgb-d sensor application for non-contact neonatal monitoring,” in *2021 IEEE Sensors Applications Symposium (SAS)*, 2021, pp. 1–6.

[19] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. Schroeder, “Learning and tracking the 3d body shape of freely moving infants from rgb-d sequences,” *IEEE Transactions on Pattern Analysis &amp; Machine Intelligence*, vol. 42, no. 10, pp. 2540–2551, 2020.

[20] N. Hesse *et al.*, “Learning an infant body model from rgb-d data for accurate full body motion analysis,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I.* Springer, 2018, pp. 792–800.

[21] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, “Learning and tracking the 3d body shape of freely moving infants from rgb-d sequences,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2540–2551, 2020.

[22] N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. S. Schroeder, “Computer vision for medical infant motion analysis: State of the art and RGB-D data set,” in *Computer Vision - ECCV 2018 Workshops.* Springer International Publishing, 2018.

[23] G. Wu, F. van der Helm, D. Veeger, M. Makhsous, P. Roy, C. Anglin, J. Nagels, A. Karduna, K. McQuade, X. Wang, F. Werner, and B. Buchholz, “Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion - part ii: Shoulder, elbow, wrist and hand,” *Journal of biomechanics*, vol. 38, pp. 981–992, 06 2005.

[24] G. Wu, F. C. van der Helm, H. (DirkJan) Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz, “Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part ii: shoulder, elbow, wrist and hand,” *Journal of Biomechanics*, vol. 38, no. 5, pp. 981–992, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002192900400301X

[25] S. van Sint Jan and P. Allard, *Color Atlas of Skeletal Landmark Definitions: Guidelines for Reproducible Manual and Virtual Palpations.* Churchill Livingstone/Elsevier, 2007.

[26] R. Neptune and M. Hull, “Accuracy assessment of methods for determining hip movement in seated cycling,” *Journal of Biomechanics*, vol. 28, no. 4, pp. 423–437, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002192909400080N

[27] R. M. Ehrig, W. R. Taylor, G. N. Duda, and M. O. Heller, “A survey of formal methods for determining the centre of rotation of ball joints,” *Journal of Biomechanics*, vol. 39, no. 15, pp. 2798–2809, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002192900500446X

[28] S. S. U. Gamage and J. Lasenby, “New least squares solutions for estimating the average centre of rotation and the axis of rotation,” *Journal of Biomechanics*, vol. 35, no. 1, pp. 87–93, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021929001001609

[29] S. J. Piazza, A. Erdemir, N. Okita, and P. R. Cavanagh, “Assessment of the functional method of hip joint center location subject to reduced range of hip motion,” *Journal of Biomechanics*, vol. 37, no. 3, pp. 349–356, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021929003002884

[30] L. Herda, P. Fua, R. Plankers, R. Boulic, and D. Thalmann, “Skeleton-based motion capture for robust reconstruction of human motion,” in *Proceedings Computer Animation 2000*, 2000, pp. 77–83.

[31] H. Kanazawa, Y. Yamada, K. Tanaka, M. Kawai, F. Niwa, K. Iwanaga, and Y. Kuniyoshi, “Open-ended movements structure sensorimotor information in early human development,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 1, p. e2209953120, 2023.

[32] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen, “Opensim: open-source software to create and analyze dynamic simulations of movement,” *IEEE transactions on biomedical engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.

[33] W. Zhao, J. Chai, and Y.-Q. Xu, “Combining marker-based mocap and rgb-d camera for acquiring high-fidelity hand motion data,” *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 33–42, 2012.

[34] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, “Deep 3d human pose estimation: A review,” *Computer Vision and Image Understanding*, vol. 210, p. 103225, 05 2021.

[35] L. Needham, M. Evans, D. P. Cosker, L. Wade, P. M. McGuigan, J. L. Bilzon, and S. L. Colyer, “The accuracy of several pose estimation methods for 3d joint centre localisation,” *Scientific reports*, vol. 11, no. 1, p. 20673, 2021.

[36] I. Van Crombrugge, S. Sels, B. Ribbens, G. Steenackers, R. Penne, and S. Vanlanduit, “Accuracy assessment of joint angles estimated from 2d and 3d camera measurements,” *Sensors*, vol. 22, no. 5, p. 1729, 2022.

[37] H. I. Shin, H.-I. Shin, M. S. Bang, D.-K. Kim, S. H. Shin, E.-K. Kim, Y.-J. Kim, E. S. Lee, S. G. Park, H. M. Ji *et al.*, “Deep learning-based quantitative analyses of spontaneous movements and their association with early neurological development in preterm infants,” *Scientific Reports*, vol. 12, no. 1, p. 3138, 2022.

[38] K. Nymoen. (2022, Oct) Infrared marker-based motion capture. Accessed: 2023-04-30. [Online]. Available: https://www.futurelearn.com/info/courses/music-moves/0/steps/12692

[39] “Miqus - refined motion capture camera |qualisys,” https://www.qualisys.com/cameras/miqus/, [Online, accessed May 2023].

[40] J. H. Challis, “A procedure for determining rigid body transformation parameters,” *Journal of Biomechanics*, vol. 28, no. 6, pp. 733–737, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002192909400116L

[41] N. Chernov. (2023) Circle fit (pratt method). Accessed: 2023-05-14. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/22643-circle-fit-pratt-method

[42] V. Pratt, “Direct least-squares fitting of algebraic surfaces,” *ACM SIGGRAPH computer graphics*, vol. 21, no. 4, pp. 145–152, 1987.

[43] P. Besl, “A method forregisttation of 3-d shapes,” *Trans. PAMI*, vol. 14, no. 2, 1992.

[44] D. G. Kyrollos, J. B. Tanner, K. Greenwood, J. Harrold, and J. R. Green, “Noncontact neonatal respiration rate estimation using machine vision,” in *2021 IEEE Sensors Applications Symposium (SAS)*, 2021, pp. 1–6.

[45] M. Contributors, “Openmmlab pose estimation toolbox and benchmark,” https://github.com/open-mmlab/mmpose, 2020.

[46] "Which intel realsense device is right for you?" https://www.intelrealsense.com/which-device-is-right-for-you/, [Online, accessed May 2023].