Dissertation zur Erlangung des Doktorgrades der
Technischen Fakultät der
Albert-Ludwigs-Universität Freiburg im Breisgau

Unobtrusive Medical Infant Motion Analysis from RGB-D Data

Nikolas Hesse

Albert-Ludwigs-Universität Freiburg im Breisgau
Technische Fakultät

Dekan

Prof. Dr. Hannah Bast

Referenten

Prof. Dr. rer.nat. Ulrich G. Hofmann

Prof. Dr. Michael J. Black, Max Planck Institute for Intelligent Systems, Tübingen, Germany

Datum der Promotion

11.07.2019

Contents

Nomenclature

3D	Three dimensional – generally refers to three dimensions of space
4D	Four dimensional – 3D over time
ACC	Accuracy
AJPE	Average joint position error
BoW	Bag of words approach
CNN	Convolutional neural network
CoM	Centroid of motion
CP	Cerebral palsy
CPU	Central processing unit
CSGM	Cramped synchronized general movement
cUS	Cranial ultrasound
DA	Definitely abnormal
DBN	Dynamic Bayes net
DTW	Dynamic time warping
FMs	Fidgety movements
FN	False negative
FP	False positive
FPS	Frames per second
GMA	General movement assessment
GMs	General movements
GPU	Graphics processing unit

HR	High risk
ICP	Iterative closest point
IMU	Inertial measurement unit
K-NN	K-nearest neighbor
LDOF	Large displacement optical flow
LOOCV	Leave-one-out cross validation
MA	Mildly abnormal
MRI	Magnetic resonance imaging
NE	Neurological examination
NO	Normal optimal
NS	Normal suboptimal
PMA	Post menstrual age
RGB	Red-green-blue – usually referring to a color image
RGB-D	Red-green-blue-depth – color and depth image
SD	standard deviation
SMIL	Skinned Multi-Infant Linear model
SMPL	Skinned Multi-Person Linear model
SVM	Support vector machine
TN	True negative
TNR	True negative rate
TP	True positive
TPR	True positive rate
WCA	Worst-case accuracy

Abstract

Cerebral palsy (CP) is the most common motor disorder in children, and is caused by injuries of the brain shortly before, during, or after birth. With traditional methods, a reliable diagnosis of CP can generally not be made until after the first year of life or even later. Based on the discovery that the condition of the nervous system of young infants is reflected in the quality of their spontaneous movements, the general movement assessment (GMA) was developed. GMA lets trained experts detect CP in infants as young as four months. However, regular practice and re-calibration are required, and GMA suffers from human variability. CP manifests itself in a variety of symptoms, which complicates finding the right parameters for predicting the risk of CP.

A cheap, automated system would allow the widespread screening of infants in order to concentrate early intervention efforts on affected children. We divide the system into the motion capture stage and the motion analysis stage. Although the extraction of motion, respectively body pose, from images and videos is a very active area of computer vision research, most of the proposed approaches focus on adults and can not be directly transferred to infants.

In this thesis, we present multiple contributions towards an unobtrusive system for medical motion analysis of infants. First, we propose an approach for estimating 3D pose from depth images, speeding up the training procedure by nearly two orders of magnitude compared to previous approaches. Second, we develop a model-based approach for markerless full-body tracking. We learn the *Skinned Multi-Infant Linear model* (SMIL) from low-quality RGB-D data and accurately register it to sequences of moving infants, capturing shape and pose while handling self-occlusions and fast movements. We show that our method captures enough motion detail for GMA. Third, we develop an approach towards predicting the GMA class from motion sequences, which is based on features of motion complexity and variation to capture characteristics of general movements.

Our methods enable accurate tracking of infant shape and pose, and enable applications like monitoring of therapy/disease progression, tracking growth or motor development, and detection of malnutrition. By publicly releasing the learned model and a synthetic, but realistic data set of moving infants, we hope to foster research focused on infants.

Zusammenfassung

Die Zerebralparese (CP) ist die häufigste frühkindliche Bewegungsstörung und wird durch Schädigungen des Gehirns vor, während, oder kurz nach der Geburt verursacht. Eine zuverlässige Diagnose kann mit traditionellen Untersuchungen meist erst gestellt werden nachdem das Kind ein Jahr oder älter ist. Ausgelöst durch die Entdeckung, dass sich der Zustand des infantilen Nervensystems in der Qualität der Spontanbewegungen widerspiegelt, wurde das General Movement Assessment (GMA) entwickelt. GMA erlaubt es geübten Experten Zerebralparesen schon bei Säuglingen unter dem Alter von vier Monaten zu erkennen. Die adequate Durchführung von GMA erfordert jedoch kontinuierliches Training sowie regelmäßige Re-Kalibrierung. Durch die menschliche Komponente wird GMA zwangsläufig von der subjektiven Wahrnehmung beeinflusst. CP äußert sich in einer Vielzahl von Symptomen, was das Finden von guten Klassifikationsparametern erschwert.

Ein kostengünstiges automatisiertes System würde flächendeckende Screenings von Säuglingen ermöglichen, um möglichst früh Therapieinterventionen auf betroffene Kinder zu fokussieren. Solch ein System kann in die beiden Bestandteile *Bewegungserfassung*, sowie *Analyse, bzw. Klassifikation* unterteilt werden. Obwohl die Erfassung von Bewegungen, bzw. Körperposen aus Bildern und Videos ein sehr aktives Forschungsfeld im Bereich Computer Vision ist, sind die meisten Methoden auf Erwachsene ausgerichtet und können nicht direkt auf Säuglinge übertragen werden.

In dieser Dissertation präsentieren wir mehrere Beiträge zu einem automatischen System für die medizinische Bewegungsanalyse von Säuglingen. Erstens entwickeln wir einen Ansatz für die 3D-Posenschätzung aus Tiefenbildern, bei dem wir den Trainingsprozess im Vergleich mit vorherigen Methoden um fast zwei Größenordnungen beschleunigen können. Zweitens führen wir einen modellbasierten Ansatz für das markerlose Ganzkörper-Tracking ein. Wir lernen das *Skinned Multi-Infant Linear model* aus niedrigqualitativen RGB-D Daten und entwickeln eine Methode zur exakten Registrierung von Modell mit Sequenzen von sich bewegenden Säuglingen. Mit diesem Ansatz können wir Form und Pose erfassen und dabei Selbstverdeckungen und schnelle Bewegungen bewältigen. Wir zeigen, dass wird damit genug Bewegunsdetails für GMA erfassen. Drittens entwickeln wir eine Methode um die GMA-Klasse anhand von den erfassten Bewegungssequenzen vorherzusagen. Dabei repräsentieren Merkmale der Bewegungskomplexität und- variation die Charakteristiken der General Movements.

Unsere Methodik bietet exaktes Tracking der Form und der Pose von Säuglingen, was weitere Anwendungen wie z.B. die Überwachung von Therapie- oder Krank-

heitsfortschritten, die Verfolgung der Größen- oder Motorikentwicklung, oder auch die Erkennung von Mangelernährung ermöglicht. Durch die öffentliche Freigabe des gelernten Modells, sowie eines synthetischen aber realistischen Datensatzes der Sequenzen sich bewegender Säuglinge enthält, hoffen wir den Fortschritt in diesem Forschungsbereich zu fördern.

1. Introduction

The diagnosis of a congenital disorder in an infant has a drastic effect on a family's life. In many cases, the chances of improving the outcome seem to be higher the earlier a disorder is detected, and families can be referred to specialized institutions [Nov+17]. This way, a head start in treatment, as well as support for the families is provided. Not all diseases can be diagnosed at an early age, but for the most common motor disorder in childhood, cerebral palsy (CP), a method for early detection exists. CP is caused by damage to the brain around the time of birth and shows different symptoms and degrees of severity. The disease causes many limitations in daily life for affected children. The severity of CP ranges from mild forms, where patients are able to walk without limitations, to severe cases, where patients lack any voluntary control of movement and have a restricted ability to maintain trunk postures.

The *general movement assessment* (GMA), which is based on the analysis of movement quality, lets trained experts predict the risk of CP at an early age. However, the number of GMA experts is limited, and the method is prone to human variability and requires regular training and re-calibration efforts to assure adequate assessment. For this reason, the automation of such a method for infant motion analysis would be beneficial. Key to an automated motion analysis[1] system is the accurate capture of motions in order to transform them to a machine-readable, numerical form.

1.1. Vision-based Motion Capture and Analysis

The field of computer vision aims at extracting high-level information about a scene from images, in our case human motion[2]. Computer vision influences many areas of everyday life, and automated methods for recovering human motion from images or videos have applications in virtual reality [Lee+02; Cha+11], robotics [Wel+17], human-computer interaction [MA07], sports [Par+17] and many more [Moe+06].

[1] The term "motion analysis" is sometimes used as a synonym for action or activity recognition, i.e., assigning a class label like "running" or "dancing" to an input sequence. In our work, we do not aim at finding the activity class for infant movements, but to assign an output value that describes the state of motor development, respectively the risk of affection with CP to an infant movement sequence.

[2] The term "motion capture" is often used to describe marker-based motion tracking methods, as is known from movies for animating artificial characters. In this thesis, we use "motion capture" to describe the extraction of motion information from images with arbitrary techniques.

In medicine, there are many use cases relying on capturing motion and/or shape from visual data: the assessment of human motion quality [Tao+15], rehabilitation using Kinect [Sin+16; Sin+17b], improving body positioning for medical scanning [Sin+17a], Alzheimer disease assessment [Iar+14], measuring performance of Parkinson's disease patients [Kuh+17], quantification of Multiple Sclerosis progression [Kon+14], or breathing analysis [Tso+14].

In this thesis, we present methods for acquiring 3D motion from data captured by a single low-cost RGB-D sensor. Based on the captured motions we target a method for motion quality assessment to predict the risk of CP in infants.

1.2. Problem Statement

This work tackles the problem of quantifying the risk of CP from motion sequences. Instead of directly predicting CP outcome, we use the general movement assessment as an intermediate step, similar to clinical practice. Machine learning techniques are consistently improving at making predictions from data. This data has to be transformed to a numerical form. We regard motion in the sense of body pose over time, i.e., body joint angles, and/or joint positions.

This leads to the second problem: acquiring body pose. State-of the art motion capture methods can not be directly transferred to infants, e.g., gold standard marker-based Vicon systems, since the capture of motions has to be performed without influencing the infant's movements. We further want to estimate the positioning and configuration of the body in 3D using a low-cost and simple sensor setup. Therefore, we utilize *RGB-D* sensors, which stands for Red Green Blue-Depth and provides color images together with a depth image, in which each pixel value represents the distance of the camera to the scene. The depth image can be transformed to a 3D point cloud using the camera parameters.

There is no definition of how much "motion detail" is necessary for medical motion analysis. We target high accuracy with a maximum of motion details while relying only on unobtrusive, low-cost hardware.

In this thesis, we aim at answering the following two research questions:

- Can we capture 3D infant motion in sufficient detail for automated motion analysis using a single RGB-D sensor?
- Can we reliably infer GMA ratings from the captured motions?

Thesis statement:

> A cheap, automated system for early-detection of CP in infants is possible.

1.3. Contributions

We develop two methods for capturing 3D motion from RGB-D sequences at different levels of detail, a method for learning a body model from RGB-D data, and a method to infer the GMA class from the captured motions. The combination of our contributions is a step towards a complete system for automated infant motion analysis. We make the model and an RGB-D data set of moving infants available to the public.

Body pose estimation in depth images using random ferns

We develop a method for infant pose estimation that is inspired by the commercial body tracking of the Microsoft Kinect sensor. We predict body part labels for each pixel of an input depth image, which lets us infer the body joint positions in 3D.

The approach used in the Kinect relies on a random decision tree classifier. Decision trees have a computationally expensive training procedure, since each node inside the tree depends on its predecessors. The training time therefore scales exponentially with tree depth.

We propose the use of a simplified variant of decision trees, namely random ferns. In ferns, the split nodes in each level of the tree are the same, which implies that the order in which the nodes are evaluated is irrelevant. This allows for much faster training, but at the cost of a reduced strength of the classifier. We compensate for that by combining many simple ferns, instead of using few complex trees. We achieve a speed up of the training procedure by nearly two orders of magnitude, from the Kinect training taking one day on a 1000 core cluster to 9 hours on a workstation with 32 cores. We further advance the accuracy of our approach by adding a feature selection step before training, making it rotation invariant, and adding kinematic constraints. However, the method shows limitations in generalizing to poses that were not included in the training data, as well as very complex poses. These limitations motivate the development of a more sophisticated model-based method.

Learning and tracking infant 3D body shape from low-quality RGB-D data

Statistical body models are generally learned from thousands of high-quality 3D scans in predefined poses. For infants, no such repository exists, which is why we learn an infant body model from low-quality, incomplete RGB-D data.

We create an initial infant model based on the adult Skinned Multi-Person Linear body model (SMPL). We fit this initial model to sequences of freely moving infants. To constrain the fitting procedure, we gather information from RGB images, depth

images, and background table information. We accumulate shape information over sequences of freely moving infants, which allows us to compensate for the incompleteness of data and to create one personalized infant shape per sequence. We learn an infant-specific shape space from all personalized shapes, and a prior over plausible poses from thousands of captured real infant poses. This constitutes our new *Skinned Multi-Linear Infant body model* (SMIL).

Experiments on more than two hours of recordings, containing more than 200K frames, show roughly 99% correct poses, and an average model to scan distance of 2.5 mm. The data contains challenging sequences with fast motion and strong self occlusions. To show the usability for medical motion analysis, we let two experts perform the general movement assessment on videos of our synthetic model registration results and, separately, on the corresponding RGB videos. We use a 1 to 10 scale, and define a threshold for agreement as the difference between ratings being $\leq$1. The first rater achieves an agreement between synthetic and real videos of over 90%, the second rater of nearly 80%. These results let us believe that our method captures enough detail for allowing GMA.

Contrary to previous approaches for infant pose estimation, which mostly estimate joint positions or rely on 2D data, we capture accurate 3D full-body shape and pose, including joint positions and joint angles. We make SMIL publicly available for research purposes.

We use SMIL to create the first realistic RGB-D data set of moving infants, which we also make available to the public.

Towards automated GMA prediction from pose sequences

We develop a method towards predicting the GMA class from captured motion sequences. We assume that sequences with similar ratings contain motions that share similar characteristics. The variety of motions that an infant generates is large, and a typically developing infant will most probably show some movements that share characteristics with movements of an infant with abnormal motor development. Over the course of a sequence, the number of "good" movements is assumed to be higher for healthy infants than that of infants with CP.

We develop features based on motion complexity and variation that capture characteristics of general movements. We train a classifier to predict whether an infant shows "definitely abnormal" movement quality, which has been shown to be a strong indicator for a high risk of CP. We conduct several baseline experiments that show the potential of the proposed features.

Additionally, we create a latent feature space to automatically group motion snippets of similar characteristics. This approach, however, does not achieve satisfactory result, which we attribute to the unconstrained nature of infant motions. A fine-grained labeling of movements with respect to GMA characteristics could produce

a representation in feature space that fully exploits the potential of the proposed features.

1.4. Ethics

Studies presented in this thesis were approved by the ethics committee of Ludwig-Maximilians-Universität München (LMU) under project number 454-16 and in compliance with the Declaration of Helsinki. All parents were informed about the recordings in a personal conversation and gave written informed consent prior to the study and could discontinue recordings or retract from the study at any time. The choice of whether or not to participate in the study was guaranteed to not have any influence on treatment. Complete patient data and recordings were safely stored at LMU. For this thesis, only a subset of the data was accessed, namely RGB-D data, age at recording, pseudonymized patient ID (for identifying multiple recordings of the same infant) and GMA ratings. From this data, no connections to patient identities could be made.

The recording procedure was fully integrated in the clinical examination protocol and did not produce any significant temporal overhead for patients, doctors or clinical staff.

1.5. Thesis Outline

In chapter 2, we give an overview of the medical background of early detection of CP and GMA. We review existing approaches towards GMA automation and analyze the underlying motion capture methods in detail. We discuss advantages and disadvantages before defining requirements for an automated motion analysis system. We present the system setup and discuss different types of RGB-D sensors as well as data properties in chapter 3. Our approach to pose estimation from depth images using random ferns is described in chapter 4. We identify limitations, and propose multiple extensions to resolve them. In our second contribution, we present methods for learning and tracking infant body shape from RGB-D sequences (chapter 5). We describe the creation of an initial infant model, which we register to the sequences, and how this enables us to learn an infant specific shape space and pose prior. We show in a case study on GMA that the proposed method captures enough motion detail for GMA. We use the learned body model for generating a data set of synthetic moving infants for evaluation of RGB-D motion capture methods in chapter 6. In chapter 7, we show how motion measurements can provide information to differentiate between multiple diseases. We present a method towards predicting the GMA class from pose sequences. To conclude, we summarize our contributions and discuss implications for future research (chapter 8).

2. Infant Motion Analysis

Human motion analysis is a valuable tool in medicine for a multitude of tasks, e.g., gait analysis [Whi96], fall risk assessment [SC+00; Per+01], or measuring Multiple Sclerosis progression [Kur83; Fis+99]. Medical motion analysis is not only applicable to adult patients, but can be used for the early detection of neurodevelopmental disorders like cerebral palsy (CP) in infants at an early age [Tac+17].

In this chapter, we provide the medical backgrounds motivating this thesis. We give an overview of causes and symptoms of CP as well as how it is diagnosed and why early detection is important, but difficult (Sec. 2.1). We describe the general movement assessment (GMA), which is the most reliable method for CP detection at a young age (Sec. 2.2). This method lets trained experts evaluate the state of the developing nervous system based on the analysis of spontaneous movements[1]. We present a review of systems aiming at automating CP detection in Sec. 2.3. After giving an overview of detection rates and types of motion features that are used for patient classification, we discuss limitations and benefits of different approaches (Sec. 2.4) leading to our definition of requirements for an automated system for infant motion analysis (Sec. 2.5).

2.1. Cerebral Palsy

Cerebral palsy is the most common neurodevelopmental disorder in children, with a prevalence of roughly 2 per 1000 live births in high-income countries [Osk+13; Sel+16]. Higher prevalence has been reported for the United States at 3 - 4 per 1000 live births [Boy+11; Chr+14].

Definition (Rosenbaum et al. [Ros+07]):

> "Cerebral palsy (CP) describes a group of permanent disorders of the development of movement and posture, causing activity limitation, that are attributed to non-progressive disturbances that occurred in the developing fetal or infant brain. The motor disorders of cerebral palsy are often accompanied by disturbances of sensation, perception, cognition, communication, and behavior, by epilepsy, and by secondary musculoskeletal problems."

[1] Please note that terms like "abnormal" refer to the movement quality shown by an infant, not the infant itself.

The most prominent impairments from CP are related to muscle tone, motor abilities and musculoskeletal problems [Ros+07]. These often are not easily noticeable in the first months of life, but become more and more apparent over time, when certain developmental milestones are not reached at the expected age [HA10]. Depending on the severity of CP, some children will not be able to walk, while others develop pathological gait patterns like tip-toeing gait, which is caused by tightness of the Achilles tendon, scissoring gait due to tightness of hip abductors, or crouching gait (hamstring spasticity). Orthopedic surgery, medication, or external fixation may help to improve such deformities [GS03]. Approximately 60% of CP patients can walk either independently or with aids as adults [Him13]. However, gait abilities decline through adulthood and the risk of falling increases [MM14].

This is in line with the description of CP being "non-progressive", i.e., the cause persists, but symptoms may change over time. This implicates that there is no cure for CP, which is why therapy aims at helping patients and their families to adapt their lives to coping with the disorder [Pal+12].

Affected children receive physical, occupational and speech therapy after a high risk of CP has been diagnosed [Nov+17]. The intervention efforts have become more oriented towards putting the emphasis on family life and increasing the independence of patients instead of solely trying to improve motor skills [Dir+11].

CP is a movement disorder, but CP patients often show additional difficulties with vision, hearing, thinking, learning, behavior and communication [Ros+07]. Although people affected by CP have a higher risk of learning disorders, they have normal intelligence [Jen+07]. However, they suffer from decreased life expectancy [Hut06].

Different *types of CP* can be identified according to the area of the brain that is damaged [Che97].

1. *Spastic CP* is the most common type, affecting roughly 70% of CP cases [Sta+00], with the most prominent impairment being hypertonia and muscle tightness (spasticity). It is caused by damages to cortical motor areas and/or white matter.
2. *Ataxic CP* is induced by injuries to cerebellar structures, and leads to decreased muscle tone and a loss of muscle coordination and balance.
3. *Dyskinetic CP* is associated with damage to the basal ganglia, and causes the inability to control muscle tone, and therefore leads to both hyper- and hypotonia.
4. A mixture of aforementioned types is also possible.

The Gross Motor Function Classification System [Pal+97] is commonly used for classifying the *level of severity* of CP. Patients are assigned to one of five classes based on the motor abilities like standing, sitting, walking and the dependency on mobility aids like wheelchairs.

Table 2.1.: Prevalence of CP in Europe in 2003 according to [Sel+16], number of affected children per 1000 live births. Different birth weight groups: normal (>2499g), moderately low (1500g - 2499g), very low (1000g - 1499g), and extremely low birth weight (<1000g).

Birth weight	Normal	Moderately low	Very low	Extremely low	Overall
Prevalence	0.89	6.2	35.9	42.4	1.77

Causes. The damage to the brain that leads to CP is mostly inflicted during pregnancy or the first months of life [HA14]. Multiple possible causes have been identified, but the occurrence of neither one necessarily causes CP. The risk of CP is highly increased in the event of critical incidents during birth, like placental abruption, cord prolapse and uterine rupture [Nel08]. However, such events are uncommon and therefore are not a main factor for CP. According to [Nel08], other causative factors for CP include prematurity, intrauterine exposure to infection or maternal fever in labor, ischemic stroke, congenital malformations, atypical intrauterine growth, and complications of multiple gestations. Prematurity, and its connection to low birth weight, has been identified by other researchers as an important risk factor which increases the probability of CP by a factor of up to 40 [Sel+16] (c.f. Tab. 2.1).

Diagnosis and early detection. Historically, the age at which CP can be reliably identified with traditional methods, like physical examination in combination with the medical history, ranges from 12 to 24 months [Pan08; Gor+09]. Naturally, the more severely an infant is affected by CP, the easier and the earlier it is detected. At young age, there are large changes in the developing brain, which may have an effect on the outcome [Kol+13]. A diagnosis therefore has to be made with caution [Pan08].

By waiting until CP can be diagnosed very accurately, a critical time window for intervention during a period of brain 'plasticity' for maximizing functional outcomes may be missed [Nov+17]. Within the first year of life, dendrites and synapses are produced at a high rate. Animal studies indicate that this period offers a good opportunity for intervention [HA14].

However, open questions remain. There has been a lack of evidence for the effect of early intervention, and it is not clear what the best type of intervention is [Spi+15; HA14; Her+15; Mor+16a]. It seems that different types of intervention are required for term-born infants and infants born preterm [BHHA05] and that intervention has a stronger effect on cognitive than on motor development [Kol+13]. The topic of early intervention is actively researched [Hie+11; Nov+17; HA+17].

Even though the effect of intervention is uncertain, it is important to identify affected children in order to refer them to specialized facilities where the development is monitored and parents are informed and supported in an optimal way [Ein+97;

Bos+13; Her+15]. Therefore, clinicians aim to detect high-risk infants as early as possible [McI+11]. New diagnostic tools like GMA allow accurate prediction of CP or "high risk of CP" before 6 months of age [Nov+17].

Metrics for CP detection. The predictive validity of medical diagnostic methods is generally reported as the *sensitivity* and *specificity*. The sensitivity is the percentage of infants who are correctly identified as being affected by CP, and the specificity is the percentage of infants who are correctly identified as not being affected by CP.

More formally, each sample of an evaluation set belongs to one of the four classes

- True positive (TP): infant affected by CP correctly classified as affected,
- False positive (FP): healthy infant incorrectly classified as affected by CP,
- True negative (TN): healthy infant correctly classified as healthy,
- False negative (FN): infant affected by CP incorrectly classified as healthy.

Sensitivity, also called the *true positive rate* (TPR) can be written as

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

with P denoting the number of "positive" cases in the data set, here infants affected by CP.

Specificity, the *true negative rate* (TNR), is then given as

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.2}$$

with N denoting the number of "negative" cases in the data set, here healthy infants.

In some cases, the *accuracy* (ACC) is given, which is defined as

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.3}$$

A problem of the accuracy measure is that in most studies on CP detection, the sets of positive and negative samples are highly imbalanced due to the small overall number of CP cases. This is not reflected in this measure, which is why for evaluation of CP detection, sensitivity and specificity should be preferred.

Tools for CP detection. A number of tools for the detection of CP exist, ranging from Magnetic Resonance Imaging (MRI) of the brain, over cranial ultrasound and neurological examinations to neuromotor assessments, like the general movement assessment (GMA). Since the focus of our work lies on GMA, we do not discuss details of the other methods and refer the interested reader to [HA14]. We present the predictive validity of the different tools in Tab. 2.2.

The general movement assessment is the most accurate tool for CP detection at an early age, with a reported sensitivity of 98% and a specificity of 91%. We give a detailed description of GMA in the next section.

Table 2.2.: Tools with predictive validity for detecting CP [Bos+13]. MRI: brain magnetic resonance imaging, cUS: cranial ultrasound, NE: neurological examination (e.g., HINE [Rom+08]), GMA: general movement assessment.

	MRI	cUS	NE	GMA
Sensitivity	86%-100%	74%	88%	98%
Specificity	89%-97%	92%	87%	91%

2.2. The General Movement Assessment

Prechtl discovered that so called *general movements* (GMs) can serve as a window to the young nervous system [Pre90]. The assessment of these general movements allows to not only describe the current state of the infant but to predict with a high probability whether or not a child will develop CP [Pre+97].

2.2.1. Motion as a Marker for Neurological Impairments

It was found that spontaneous movements are motor activities that are endogenously generated, opposed to reflexes, which are triggered by external stimuli [Pre90]. Animal experiments have shown that these general movements are more prone to impairments of the nervous system than reflexes [Pre90].

Prechtl [Pre90] defines general movements as

- gross movements involving the whole body,
- lasting between few seconds and one minute,
- variable sequences of arm, leg, neck and trunk movements,
- varying in intensity, force and speed, with gradual onset and end,
- mostly complex extension or flexion of arms and legs, with superimposed rotations and changes in movement direction, and
- fluent, elegant movements, creating the impression of complexity and variability.

If the quality of GMs deviates from the above definition, this is a reliable indicator of brain dysfunction [Ein+97].

GMs are present from roughly 28 weeks post menstrual age (PMA) (fetus) until 14-18 weeks post term and change their **characteristics** over time [HA04]:

- *Preterm GMs* occur between 28 weeks and 36-38 weeks PMA, consisting of "extremely variable movements, including many pelvic tilts and trunk movements".
- *Writhing GMs* are more forceful, but slower than preterm GMs and pelvis and trunk are less involved. They occur between 36-38 and 46-52 weeks PMA.

Table 2.3.: Classification system according to Hadders-Algra [HA+04]

Classification	Complexity	Variation	Fluency
Normal optimal GMs	+++	+++	+
Normal suboptimal GMs	++	++	-
Mildly abnormal GMs	+	+	-
Definitely abnormal GMs	-	-	-

- *Fidgety GMs* are shown between 46-52 and 54-58 weeks PMA. They are characterized by a continuous flow of small and elegant movements, involving all parts of the body.

Abnormal GMs until 2 months post term show one or more of the following characteristics [Pre+97; Ein+97]:

- *poor repertoire GMs* are characterized by monotonous movements and a lack of movement complexity,
- *cramped-synchronized GMs* appear rigid, with reduced smoothness or fluency, and a simultaneous contraction and relaxation of all limb and trunk muscles,
- *chaotic GMs* have mostly large amplitude or seem very abrupt, and lack fluency or smoothness and occur in chaotic order.

At fidgety age (up to 20 weeks post term), general movements are classified as abnormal if they are either *absent*, i.e., the infant never showed fidgety movements (FMs) between the age of 6 and 20 weeks post term, or *abnormal*, when amplitude, speed and jerkiness deviate from the range of normal FMs [Ein+97]. The method of GMA is very reliable, and if abnormal movement patterns persist over time the reliability increases even more [Ein+97].

Hadders-Algra proposes a different GMA classification scheme, more focused on overall complexity and variation than on fidgety movements [HA+04; HA04]. In their studies, they observed fidgety movements lacking fluency, but having sufficient complexity and variation, which they classify as *mildly abnormal* GMs. They conclude that "fluency is a feature that is easily disturbed", which is why they do not consider it as a criterion for abnormality. Mildly abnormal GMs are associated with an increased risk of *minor neurological dysfunction* (MND), attention problems and aggressive behavior at school age [HA+04]. GMs classified as *definitely abnormal* indicate a high risk of CP. Normal GMs are subdivided into normal optimal GMs, showing complexity, variation and fluency, and normal suboptimal GMs, which show less, but still sufficient complexity and variation, but lack smoothness. An overview of the classification scheme is shown in Tab. 2.3. The intra- as well as the inter-observer agreement of skilled observers is reported to be high [HA+04].

A recent study suggests that Prechtl's GMA has a higher reliability than the GMA according to Hadders-Algra [Kwo+18]. However, this review includes results for

Hadders-Algra's GMA from a study on the reliability of GMA in the general population, which is reported to be much lower than in high-risk populations [Bou+10]. The studies using Prechtl's GMA are only conducted on high-risk populations. This has led to discussions among researchers, and the strong recommendation of considering the population when applying GMA [HAP18]. On high-risk populations, the reliability of both GMA variants is high [HA+04; Phi+14; EP05].

2.2.2. Methodology of the General Movement Assessment

The general movement assessment is performed by a trained expert based on visual Gestalt perception [Pre90]. Video recordings should be favored over direct observation, to avoid distraction and to have the possibility of repeated playback at different speeds and to store the recordings for later reference [Pre90; Ein+97].

The infant should be placed in supine position and wear light clothing. It should be in a state of active wakefulness[2]. GMA is not applicable during non-nutritive sucking, hiccup, fussing, crying, or interaction with toys or parents [Ein+97].

Up to term age, duration of infant recordings should be one hour, which is reduced by selecting the best GMs in a post-processing step [Ein+97]. After term age, infants should be recorded for 5 - 10 minutes, with an absolute minimum of three minutes. Experienced observers require not more than one to three minutes of recorded GMs [EP05]. A developmental trajectory should be preferred over assessment of a single recording [EP05].

New technologies have made it easier to record infants in a relaxed atmosphere at home using a smartphone and have the recordings transferred electronically to the GMA experts [Spi+16].

The advantages of GMA include its non-intrusiveness to the infant, compared to imaging techniques like MRI. The only required equipment is a camera and a playback device, e.g., a smartphone. The method has shown to achieve high reliability and validity in predicting CP. It can be learned by anyone and does not require previous medical knowledge.

However, there are certain drawbacks. While the basic principles of GMA can be learned in two days, it takes further practice to become skilled [HA+04]. Besides from regular practice, a repeated re-calibration with gold-standard videos is required to avoid bias [Ein+97]. Although GMA has shown to have high reliability, the inter- and intra-observer agreement decreases in clinical practice [Ber+11]. GMA is dependent on the skill of the rater, and therefore is subject to human variability. The number of training courses is small – the official website of the *General Movement Trust*[3] lists five courses between November 2018 and May 2019 worldwide.

[2] This corresponds to Prechtl state 4 [Pre74].

[3] `http://www.general-movements-trust.info/47/dates`, accessed November 2018.

An automated system for GMA/CP detection could contribute to widespread screening and identification of infants in need of special treatment. In 1990, Prechtl [Pre90], referring to the technique underlying GMA, stated:

> "Gestalt perception is a powerful instrument in the analysis of complex phenomena and an instrument which cannot be replaced by automated quantification."

This viewpoint, however, could not keep researchers from working on methods for the automation of early detection of CP.

2.3. Related Work – Towards Automated Detection of Cerebral Palsy

Fueled by recent advances in computer vision, research efforts have been undertaken towards automated motion analysis for a variety of medical purposes, e.g., for patient monitoring [Ach+16], quantifying therapy or disease progression [Kon+14], or performance assessment [Cla+12]. In order to perform an automated quantitative analysis of movements, these movements have to be captured and converted to a numerical form, e.g., joint positions or angles.

We review systems aiming at the automated prediction of cerebral palsy in infants based on the assessment of motions. Although this problem is approached in different ways, the pipeline is similar for most systems, and can be divided into motion capture and motion analysis (Fig. 2.1). Motion features are extracted from captured movements, and used for training a classifier to predict the outcome. We provide an overview of prediction accuracy reported in the reviewed systems in Tab. 2.4.

Parts of this section were published as [Hes+19a].

2.3.1. Wearable Motion Sensors

Before focusing on vision-based approaches, we review systems using wearable motion sensors to capture infant motion for CP detection. Prior to the introduction of low-cost wireless technology, researchers investigated the use of wired motion sensors.

Karch et al. propose an electro-magnetical tracking system, which measures positions of electrodes attached to the infant's limbs [Kar+08]. The setup of the system is described in more detail in [Kar11]. The system provides a high spatial resolution, but the capture space is very limited (16 cm × 40 cm × 40 cm). In addition, external electro-magnetic influences, e.g., by electrical devices or wires, have to be minimized to ensure accuracy. Four electrodes are attached to the upper limb and four to the lower limb. Due to technical restrictions, only one body side is captured at a

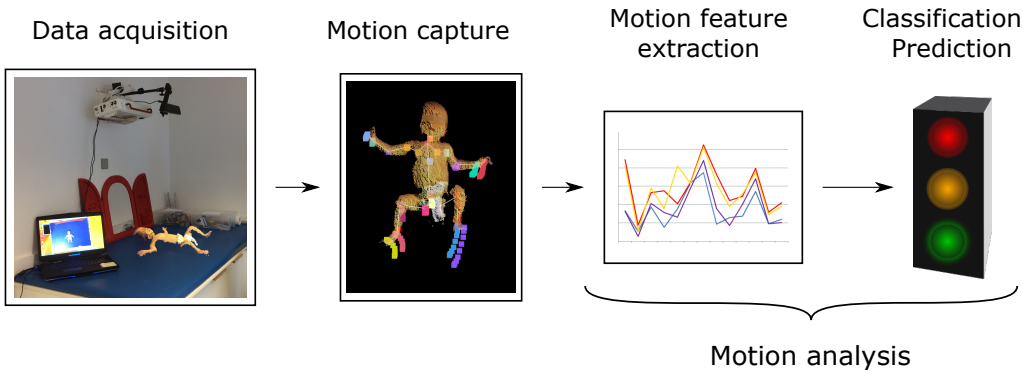


Figure 2.1.: Standard motion analysis pipeline. Motions are captured from the acquired data. Motion features are extracted and used to classify or predict the medical outcome.

time. After doing an initial calibration by performing several predefined movements with the infant's limbs, spontaneous movements are captured. These are mapped to a biomechanical model to extract joint positions and angles. Based on these measurements, the authors develop a complexity score to distinguish between a pathological group and a control group [Kar+10]. They train a principal component model from manually selected complex movements of five healthy infants. The complexity score for new movements is determined by calculating the accordance to the model. In further work, they develop a stereotypy score, which describes movement sequences with respect to repetitiveness and self-similarity, which are indicators for neurological deficits [Kar+12; Phi+14]. In two studies they achieve a sensitivity of 90% and a specificity of 95 - 96% for correctly predicting CP.

The technical approach is developed in a coherent and thoughtful way. However, the limitations of the recording system seem to be too substantial to allow its use in standard clinical practice. Tedious setup and calibration procedure, together with easily disturbed measuring procedure suggest the use of different sensors.

Researchers have found inertial measurement units (IMUs) to be a better choice for measuring infant motion. IMUs measure rotational forces, angular rates and velocities using a combination of accelerometers, gyroscopes, and sometimes magnetometers.

One of the first studies on using accelerometers for assessing infant motor development was presented by Ohgi et al. [Ohg+08]. They attach a single accelerometer to one wrist of 7 infants with brain injuries and 7 healthy infants to compare the acceleration time series. They found that movements of infants with brain injuries are more chaotic and disorganized.

Heinze et al. attach one wired accelerometer to each extremity to extract acceleration and velocity over time [Hei+10]. Accelerometers are subject to measurement bias/noise. When the signal is temporally integrated to obtain velocities or posi-

Table 2.4.: Overview of automated CP detection approaches. We present the number of subjects, type of motion features used for classification, and prediction results. *HR*: infants at high risk of CP, *CP*: infants with outcome CP. Results from multiple publications of the same first author are summarized in one row. Studies containing only HR subjects predicted the label HR instead of CP.

Publication	# infants	Features	Sensitivity | specificity	Accuracy
Meinecke [Mei+06]	22 (7 HR)	8 motion features (lower limbs)	100% | 70%	73%
Adde [Add+09; Add+10; Add+13]	82 (32 HR), 30 (13 CP), 52 (9 CP)	Centroid of motion	81.5, 85, 89% | 70, 71, 74%	-
Stahl [Sta+12]	82 (15 CP)	Frequency + motion distance	85.3% | 95.5%	93.7%
Rahmati [Rah+14a; Rah+15a; Rah+16]	78 (14 CP)	3 features from [Mei+06], frequency	50, 92, 86% | 95, 87, 92%	87, 88, 91%
Kanemaru [Kan+14]	145 (16 CP)	Jerk index	Higher jerk index in CP	-
Heinze [Hei+10]	23 (4 CP)	32 features from [Mei+06]	100% | 83-89%	88-92%
Karch [Kar+12]	62 (10 CP)	Stereotypy score	90% | 96%	-
Philippi [Phi+14]	67 (10 CP)	Stereotypy score	90% | 95%	-
Orlandi [Orl+18]	127 (16 CP)	Many motion features	44% | 99%	92%

tions, the calculated values drift linearly for velocities or quadratically for positions over time. This is why an additional noise reduction method is generally applied to achieve improved accuracy. The authors refrain from calculating 3D trajectories from the recorded data due to imprecise and unstable results. A subset of parameters proposed in a different approach [Mei+06], related to acceleration and velocity, is selected. A sensitivity of 100% at a specificity of 83% is achieved for a group of 17 healthy and 4 high risk infants. However, to increase the number of items in the data set, multiple recordings are performed in the first five months of life for each infant and treated these recordings as statistically independent. This way, the overall number of used recordings is 139 (107 healthy, 32 pathologic) for training and 70 (53 healthy, 17 pathologic) for evaluation. The percentage of correctly identified pathological motion sequences varies between 50 and 71%, depending on the age at the time of measurement, at 100% correctly identified healthy sequences.

Technological progress has led to a decreasing size and cost of wearables and better usability by the introduction of wireless sensors. Instead of aiming at the direct prediction of CP, some researchers try to identify so called cramped-synchronized general movements (CSGMs, cf. Sec. 2.2.1), which are a sign of abnormal motor behavior.

Gravem et al. use lightweight wireless accelerometers, which they attach to head and limbs of ten infants that were born preterm [Gra+12]. They evaluate three different classifiers that were trained on manually annotated data. In the text, an accuracy of 70 - 90%, average sensitivity of 99.2%, and average specificity of 99.6% is reported. However, the authors provide more detailed results in a table, where the average sensitivity is reported as ranging between 6.9% and 49.8%, and an average specificity between 76.4% and 96.4%. We do not know how the numbers given in the text are calculated. The percentages in the table seem more plausible, given the imbalance of CSGMs versus other types of movements in the data set.

Singh et al. [SP10] follow a similar approach. They use a leave-one out cross-validation scheme, and report an accuracy of 92.7% for a Support Vector Machine classifier (SVM) and 89.8% for decision trees. However, since there are many more "normal" movements (640K) than CSGMs (28K), and the goal is to detect CSGMs, accuracy seems like an unsuitable metric. In the evaluation of an SVM classifier, roughly 2K of 28K CSGMs are correctly identified (true positives), while over 26K real CSGMs are classified as normal movements (false negatives) and over 22.5K normal movements are falsely classified as CSGMs (false positives). From these numbers, we calculate a sensitivity of 7% and a specificity of 96.5%. A classifier using a Dynamic Bayes Net with a Random Forest correctly classified CSGMs in 50% of the cases (14K CSGMs), but at the same time nearly 200K normal movements were classified as abnormal. From these numbers we compute a sensitivity of 50% and a specificity of 70%. Additionally, an accuracy of 95% is reported for a baseline method that always predicts the most popular class ("normal"). Given that the proposed approaches are not able to improve over this simple baseline, the system doesn't seem useful for clinical applications. This further highlights the use of an appropriate evaluation metric.

This work is extended by Fan et al. by including temporal modeling of CSGMs [Fan+12]. On the same data set, they report 72% sensitivity and 57% specificity.

The approaches for CSGM detection seem far away from results that would make them usable in clinical practice. One problem is the low number of CSGM occurrences compared to other types of movements.

Although a recent study shows that wearable sensors do not seem to affect the leg movement frequency [Jia+18], they supposedly have a negative influence on the infant's state of happiness. Karch et al. report that recordings for two thirds of participating infants had to be stopped after re-positioning the attached sensors due to crying (and technical difficulties) [Kar+12]. Furthermore, approaches relying on attached sensors generally suffer from practical limitations like time consuming

human intervention for setup and calibration, and add the risk of affecting the movements of infants. Without proper calibration, inaccurate positioning of sensors may lead to biased results [See+14].

2.3.2. Video-based Approaches

Cameras, opposed to motion sensors, generally require no calibration, are cheap, easy to use, and can be easily integrated into standard examinations while not influencing infants' movements. This makes them more suitable for use in clinical environments, doctor's offices or even at home. Other than sensor-based approaches, vision-based approaches do not measure motions directly. More or less sophisticated methods are needed to extract motion information, e.g., by estimating the pose in every image of a video. We describe the methods for motion capture used in the current state-of-the-art in infant motion analysis, as well as the evaluation protocols for these methods. Our findings further support the need for a standardized, realistic, and challenging benchmark data set.

We review approaches that process RGB (or infrared) images for the capture of infant motion. We include methods relying on attached markers, despite posing some of the same challenges as wearable sensors. These require human intervention for marker attachment, calibration procedures and most of all possibly affect the infants' behavior or content. Still, they use computer vision for tracking the pose of the infants. We give an overview of methods used for capturing motions from video and the respective evaluation procedures in Tab. 2.5.

One of the first automated approaches for CP detection was presented in 2006 by Meinecke et al. [Mei+06]. They use a Vicon system consisting of 7 infrared cameras surrounding the infant, equipped with 20 markers. The marker positions are specified on a biomechanical model, which allows to infer the 3D positions of the markers from the recorded images. These positions can be accurately measured at a reported error of 2 mm for a measurement volume of 2 m^3. However, the system suffers from certain limitations. Due to the unconstrained movements of the infants close to the underground, the markers of the upper extremities are often occluded and therefore invisible to the cameras. The attachment of additional markers exceeded the system's capabilities and therefore, motion features of upper extremities were excluded. In order to classify the risk of CP from the recorded motions, they extract 53 motion parameters and use cluster analysis to find the eight most relevant ones for predicting CP. The system is trained using motion parameters from 4 healthy and 4 at-risk infants. The test set consists of 14 children, who are classified with an overall correct detection rate of 73% (sensitivity 100%, specificity 70%). The high cost of the system, the complex setup and calibration, and the occlusion problems stand against the highly accurate tracking of joints in 3D. This seminal work sparked research on automated CP detection, and at the same time demonstrated the challenges associated with this task.

Table 2.5.: Summary of motion capture methods and corresponding evaluation of video-based approaches for medical infant motion analysis.

First author and reference	Sensor system	Method | tracked limbs	Groundtruth generation | Reported avg. accuracy
Meinecke [Mei+06]	Vicon	7 IR cameras, 20 markers | head, trunk rot., 3D joint pos.	Vicon precision 2 mm (for measuring volume 2 m^3)
Adde [Add+09]	RGB	Difference image | centroid of motion	No evaluation
Stahl [Sta+12]	RGB	Optical flow | regular grid	Manual annotation | exemplary plot (160 frames)
Kanemaru [Kan+13]	2D marker tracking	Marker tracking | arms, shoulders, hips, legs	No evaluation
Rahmati [Rah+15b]	RGB	Optical flow | head, torso, arms, legs	Manual annotation (20 frames), Freiburg-Berkeley data set (moving objects) | < 5 cm (extracted from plot)
Machireddy [Mac+17]	RGB + IMUs	Marker tracking + IMUs | 3D pos. of limbs + chest	Human arm model on drill, with marker and IMU | plot for visual comparison of trajectories
Orlandi [Orl+18]	RGB	Optical flow | centroid, silhouette	Manual annotation | Correlation 0.83 (high)

Disselhorst et al. [DK+12] use a similar system with the same parameters. Instead of applying it for CP detection, they aim to quantify movements of different age groups for infants up to 6 months.

The high cost and effort required to use the Vicon system, together with reported problems led researchers to investigate simpler and cheaper sensor systems. A body of work considers the use of standard **2D video** cameras, which we review next.

Adde et al. propose a holistic approach [Add+09; Add+10; Add+13; Add+18] on motion tracking. Instead of estimating locations of individual limbs, they calculate the difference image between two consecutive frames to generate what they call a motion image. They calculate the centroid of motion, which is the center of the pixel positions forming the motion regions in the motion image. They construct a "motiongram" by compressing all motion images of a sequence either horizontally or vertically by summing over columns, respectively rows, and stacking them to give a compact impression on how much an infant moved, and where the movements happened. As features for classification, they use the quantity of motion, which they define as the percentage of motion pixels in a motion image, the standard deviation (SD) of the centroid of motion (CoM), as well as the SD of velocity and the SD of acceleration of the centroid of motion. They perform multiple studies

based on these features.

Instead of directly predicting CP outcome, motions are either classified as fidgety (FM) or non-fidgety movements in 137 recordings of 82 infants, of which 27 recordings have absent FMs, and 110 contain observable FMs. A sensitivity of 81.5% and a specificity of 70% is achieved. The threshold for assigning a class to an infant recording is manually chosen after feature calculation to give the maximum accuracy for the data set.

In a follow-up study on 30 high-risk infants of which 13 were later diagnosed with CP, a sensitivity of 85% and specificity of 71% is reported for CP detection [Add+10]. The classification thresholds are again adapted to the data set and therefore take different values than in previous work [Add+09]. For this reason, a generalization to new samples for which the outcome is unknown is questionable. Splitting the data in training and test set would have produced more meaningful results.

Another follow-up study shows that using multiple recordings of the same infant – instead of just one – leads to an increase in prediction accuracy [Add+13]. CP detection in 52 infants obtains a sensitivity of 89% and a specificity of 74%, but the thresholds are adapted to the data set again. Motion feature take significantly different values for infants at fidgety age compared to infants at writhing age [Add+18].

From the same research group, Stoen et al. [Stø+17] conduct a study on a larger population containing 241 recordings of 150 infants at an age of 10 to 15 weeks. Motion features are applied to distinguish between infants with normal FMs and infants with sporadic or absent FMs. By choosing fixed values for sensitivity of 80% and specificity of 80%, respectively 90%, they report a referral rate to GMA of 27% at a false negative rate of 20%, respectively a referral rate of 40% at a false negative rate of 10%.

Other researchers have used more sophisticated methods for motion extraction. Stahl et al. use a motion tracking method based on optical flow between consecutive RGB frames [Sta+12]. Points on a regular grid distributed across the image are tracked over time. To evaluate the motion tracking accuracy, five points are manually selected from the grid as head, hands and feet, and the estimated positions are compared with manually annotated ground truth in a plot over 160 frames. For classifying CP, motion features based on frequency, wavelet coefficients, and absolute motion distance are used to train an SVM classifier based on the statistics over the whole grid. They achieve an accuracy of 93.7 % at a sensitivity of 85.3 % and a specificity of 95.5 % with 10-fold cross-validation on a data set containing 82 infants of which 15 are affected with CP. They report that the approach loses track when fast motions occur and therefore re-initialize the tracking every 21 seconds.

Orlandi et al. use large displacement optical flow (LDOF) for tracking infant motion [Orl+18]. The user is required to manually select skin regions from multiple frames, which are used to build an appearance model for segmenting the infant from the scene. They calculate the LDOF only inside the infant silhouette to extract velocity components in both spatial directions. For evaluation of their method, they

manually annotate 15 joint positions every 10 frames in their video and compare the velocities of the annotated points to their estimated velocities. They report their estimates as being comparable and highly correlated to the manual annotations. For each video, 643 numerical features from literature on automated GMA are extracted. A feature selection method identifies the nine best features to predict typical vs. atypical movements and CP. The data set consists of 127 videos. 29 recordings are classified as containing atypical GMs, and 16 of these 29 infants were later diagnosed with CP. Their best reported results for typical vs. atypical GM classification are achieved by an AdaBoost classifier at a sensitivity of 55%, a specificity of 95%, and a resulting accuracy of 86%. The prediction of CP results in a sensitivity of 44%, a specificity of 99%, and an accuracy of 92% for a Random Forest classifier. Manual clinical GMA as a predictor of CP on their data set gives a sensitivity of 81%, specificity of 86%, and accuracy of 85%.

Rahmati et al. evolve their approach in multiple publications. Limbs and head are tracked in RGB videos based on shape and appearance [Rah+12]. This approach is extended by tracking a relatively dense optical flow field over the video [Rah+14a]. There is no quantitative evaluation of the tracking accuracy, and the authors display five, respectively eight qualitative result frames. Each point in the field produces a trajectory over time, and similar trajectories are assigned to body parts based on initial manual labeling. They classify 78 infants with a reported accuracy of 87% using a subset of three motion features proposed by Meinecke et al. [Mei+06]. The relatively high accuracy stems from the high number of non-CP subjects, the sensitivity for correctly detecting an infant which is affected by CP is just 50%, while the sensitivity is 95%. In subsequent work, they integrate prior knowledge in the form of manual segmentation of a small number of frames into their motion segmentation algorithm [Rah+14b; Rah+15b]. They evaluate the accuracy on 20 manually annotated frames from 10 infant sequences, reporting an F-measure of 96% by calculating the overlap between ground truth and estimated segmentation. They compare their tracking method to different state-of-the-art trackers on the same data set, with their tracker showing superior results. Furthermore, they evaluate their segmentation method on the Freiburg-Berkeley data set containing moving objects, e.g., cats and ducks, and compare results to an optical flow method. Their method achieves best results at an F-measure of 77%. For CP detection, they rely on features based on motion frequency [Rah+15a; Rah+16], and achieve a maximum accuracy of 92% for their video-based method, at a sensitivity of 77% and specificity of 95%, using the same data set as in previous publications.

Khan et al. estimate body parts from RGB images based on deformable appearance models, while simultaneously encoding spatial relations [Kha+18]. A manual labeling of images with 14 body parts is used to train the appearance models. Based on estimated body parts, joint positions and corresponding 2D angles are calculated. Their RGB-based approach is compared to results from our approach for 3D pose estimation from depth images [Hes+15], which is presented in chapter 4. They report an average joint position error of 1.27 cm, with worst results for head (2.03 cm) and

best for right shoulder (1.1 cm), compared to 4.1 cm of our baseline method. They further evaluate the worst-case accuracy (WCA), which denotes the percentage of frames for which all joint errors lie below a given threshold. For a threshold of 5 cm, they achieve a WCA of 95.8%, and 86.3% for a threshold of 3 cm, compared to the extended version of our approach ([Hes+17a] and Sec. 4.3) at a WCA of 90% and 85%. However, their approach is evaluated on a different data set, which makes the comparison hard to interpret. They do not explain how they calculate 3D from their estimated 2D joint positions in RGB images, so we can only assume they are projecting their 2D estimates onto the depth image. An evaluation of angles calculated from the estimated joint positions with respect to ground truth annotations results in a mean absolute error of 3 degrees. A discussion of failure cases is limited to the statement that significant occlusions pose problems to the approach. They suggest the use of multiple cameras to handle this scenario.

To overcome problems of tracking motion solely based on appearance, some researchers have attached **markers** to facilitate limb tracking in 2D videos.

Kanemaru et al. use a commercial marker tracking system (Frame-DIAS) to record 2D positions of markers on arms, shoulders, hips and legs at 30 Hz using a single camera [Kan+13]. Infants who show abrupt and synchronized movements of the limbs together with reduced activity seem to be correlated with developmental delay at the age of three years. In a subsequent study containing 145 infants of which 16 have outcome CP, they report a higher "jerk index", describing the jerkiness of movements in infants affected by CP, compared to healthy infants.

Hybrid systems using a combination of video and IMU sensors for 3D adult pose estimation have proven to produce accurate results [Mar+16; Mar+18].

Machireddy et al. [Mac+17] present a hybrid system for infants, which combines color-based marker tracking in video with IMU measurements. The different sensor types are intended to compensate for each others limitations. The IMU sensors are attached to the infant's limbs and chest, together with colored patches. The 2D positions of patches are tracked based on color thresholds. From the known patch size and the camera calibration, an estimate for the 3D position of each patch is calculated. Ground truth for evaluation of motion tracking is generated by rotating a plywood model of a human arm using a drill, equipped with one marker and one IMU. As result, the authors present plots of ground truth positions and estimated positions for a circular and a spiral motion, without providing exact numbers on accuracy. Instead of classifying CP directly, they take an intermediate step and aim at detecting *fidgety movements* (FM). They use FM and non-FM annotations from one video to train an SVM classifier with 10-fold cross-validation to distinguish between these two classes, and report an accuracy of 84%.

To conclude, Meinecke et al. showed that gold-standard motion capture methods for adults can not be easily applied to infants. Video-based approaches lack 3D information, and are subject to lighting conditions and appearance. Sensors that provide 3D information in addition to color images may resolve these problems.

2.3.3. Approaches based on RGB-D Sensors

RGB-D stands for Red Green Blue-Depth – such devices usually have a color camera and a sensor producing depth images, i.e., for every pixel of the image, the distance of the camera to the scene is provided. The depth values can be transformed to 3D coordinates using the calibration parameters of the device. Calibration needs to be performed only once per camera, and most sensor manufacturers provide either a method for transforming depth images to point clouds, or the calibration parameters. In Sec. 3.2, we review depth sensing technologies and provide examples of different sensors.

With the introduction of low-cost RGB-D sensors, motion analysis approaches started taking advantage of depth information. The probably most well-known RGB-D camera is the Microsoft Kinect, which was introduced as a gesture control device for the gaming console XBox, but soon became widely used in research due to its affordable price. The motion tracking provided by the Kinect SDK has been used for motion analysis purposes [Cla+12], but does not work for infants as it was purposed for gaming scenarios of standing humans taller than one meter. We review approaches that aim at estimating infants' poses from RGB-D data and turn our attention to the respective evaluation procedures.

Most of the previously reviewed publications provided an evaluation of CP classification accuracy. The approaches below do not exclusively focus on CP prediction, but on methods for capturing spontaneous infant movements. A summary is displayed in Tab. 2.6. Since motion capture is fundamental for an automated motion analysis system, we review all of these infant-specific systems here, and provide a review of the state of the art in general (adult) motion tracking in later chapters.

Olsen et al. transfer an approach to infants that was originally introduced for adults [Ols+14]. It is based on the assumption that extremities have a maximum geodesic distance to the body center. In a first step, the body center is localized by color filtering for the infant's clothing. Then, five points on the body which are farthest from the center are searched, i.e., the geodesic extrema. Each of them is assigned to one of classes head, left/right hand, left/right foot, based on the spatial location and the orientation of the path to the body center. Intermediate body parts like elbows, knees and chest are calculated based on fractional distances on the shortest path from body center to extremities, resulting in 3D positions of eleven joints. For evaluation, an unspecified number of frames is manually annotated with 3D joint positions. Annotated joints lie in the interior of the body, while the estimated joints lie on the body surface. Results are presented in a plot, numbers given here are read off this plot. The average joint position error is roughly 9 cm. Highest errors occur for hands and elbows (15 cm), lowest for body center, chest, and head (3 cm).

In subsequent work, the same authors assemble an infant body model from simplistic shapes (cylinders, sphere, ellipsoid) and use it to track eleven underlying body joints in depth images [Ols+15]. After determining size parameters of the body parts,

Table 2.6.: Summary of motion capture methods and corresponding evaluation of depth-based approaches for medical infant motion analysis. SD denotes standard deviation.

First author and reference	Method | tracked limbs	Ground truth (GT) generation | Reported avg. accuracy
Olsen [Ols+14]	Geodesic distances | 11 3D joint positions	Manual annotation (number of frames not specified) | 9 cm (extracted from plot)
Olsen [Ols+15]	Model-based tracking | 11 3D joint positions	Manual annotation (number of frames not specified) | 5 cm (SD: 3 cm) (extracted from plot)
Serrano [Ser+16]	Model based tracking | angles of hip, knee, and ankle	Robot leg kicking, angle comparison for knee and ankle, 250 frames | 2 - 2.5 degree error
Cenci [Cen+17]	Movement blobs | arms and legs	No evaluation
Shivakumar [Shi+17]	Optical flow + color-based segmentation | 3D positions of head, torso, hands, feet	Manual annotation of 60 frames | 8.21 cm, SD: 8.75 cm

their previous method is used for finding an initial pose. Subsequently, they use the iterative closest points algorithm to adjust the model parameters to explain the input point cloud that is calculated from the depth image. Similar to previous work, they evaluate the accuracy of their system with manually annotated 3D joint positions of an unspecified number of frames. The numbers given here are extracted from presented plots. An average joint position error of 5 cm (standard deviation (SD) 3 cm) is achieved. Largest errors occur for right hand (7 cm) and stomach (6 cm).

Opposed to previous approaches, which rely on readily available depth sensors, Shivakumar et al. introduce a stereo camera system, providing higher depth resolution than existing sensors [Shi+17]. However, passive stereo camera systems, i.e., systems without illuminating the scene with a structured light source, have a hard time finding correspondences for stereo matching in untextured parts of the scene. It is questionable whether the proposed system leads to higher quality depth images than existing active stereo, structured light, or time-of-flight sensors. They compare the estimated depth of a flat table that is covered with a textured blanket with manual measurement and report an error of 1.5 cm between measurement and average estimated depth at a distance to the camera of 55 cm. To locate the infant body center, they apply a color threshold and fit an ellipse to the colored region and track it over time. In addition to the torso center, hands, legs and head regions are selected by the user, which are then tracked based on their color. The positions of limbs are defined as the pixel in the corresponding limb region that is farthest from the body center. In case of overlap of multiple limb regions, a recovery step

distinguishes them. An optical flow method is used for estimating the motion of the limb positions in the successive frame. An evaluation is presented on 60 manually annotated frames from three sequences, showing an average error of 8.21 cm (SD: 8.75 cm) over all limbs.

The literature on infant motion tracking using RGB-D sensors is limited, which is why we also review approaches that capture only a subset of body joints, or use a different representation of body motion.

Serrano et al. limit their tracking method to lower extremities using a leg model [Ser+16]. Their approach is semi-automatic and requires manual intervention. The infant's belly is manually located from the point cloud and the tracker's view is restricted to one leg. After the length and width of each segment of the leg model are defined, the model parameters (angles) are optimized using robust point set registration. They generate ground truth for 250 frames using a robotic leg that simulates kicking movements of infants. The average angle error of the proposed method is reported with 2.5 degrees for the knee and 2 degrees for the ankle.

Cenci et al. do not use an explicit representation of spatial changes of body parts [Cen+17]. They detect motion by calculating the difference image between two depth images and applying a threshold. After processing a complete sequence, they get a map, showing the locations in which movement occurred. They perform a clustering, and assign each of the clusters to a body part, based on the assumed locations (upper left is assumed to be left arm, lower left to be left leg, etc.). There is no evaluation of the correctness of assigning blobs to limb classes. One would assume that when limbs interact with each other, or a limb is moved across the body midline, the assignment of a point to a body part will be most probably wrong. To model characteristics of spontaneous infant movements, they create an encoding scheme, in which each limb can take the state "in movement" or "not in movement". They encode all possible states of combinations to describe each motion sequence as a series of transition states, and calculate statistics over the different states.

Others. Marschick et al. propose a setup combining all kinds of sensors, with the goal to create an age-specific "fingerprint" model for different neurological conditions of interest [Mar+17a]. They propose a recording setup comprised of two video cameras, two RGB-D sensors, a microphone, wearable motion sensors, and a pressure mat. They have not yet developed methods for tracking movements from the sensors, but aim to fuse information from all modalities.

2.4. Discussion

Although promising results have been published, the problem of automated CP detection is not solved.

A common limitation of the reviewed studies is the low number of subjects affected by CP, which makes it difficult to draw conclusions about the general validity. Although being the most common motor disorder in childhood, CP is still a relatively rare disease and absolute numbers are small – at most 16 CP patients are included in the studies. The confirmed outcome is established years later, which limits the speed of creating additional data. This leads to a delay, e.g., from the introduction of new sensor technologies to the time infants can be evaluated regarding to CP outcome with these new sensors. Different CP symptoms have different characteristics, which is why results on small data sets should be taken with caution. The generalization of the results to the general population seems to be questionable [Bou+10], and most studies only include high-risk subjects.

In conclusion, the proposed methods for infant motion tracking are rather simplistic when compared to the state of the art in general (adult) motion tracking, which we review in later chapters (Sec. 4.1 and Sec. 5.2). We believe the main reason to be the small amount of (publicly) available training data. Most adult approaches rely on large sets of manually annotated or synthetically generated data to train deep neural networks or use other machine learning techniques.

Recording infants poses several challenges, which makes it laborious to generate large amounts of data. The overall number of infants compared to the number of adults is small, which makes it harder to recruit subjects. Popular adult motion data sets contain sequences of people performing previously defined actions with attached markers for ground truth generation, e.g., the CMU motion capture data set [GS01], or HumanEva [Sig+09], but infants are not instructable. A different way to approach data set creation is to leverage public sources of image or video data, like YouTube, and annotate the desired information afterwards, similar to the MPII Human Pose data set [And+14]. The time-consuming and tedious manual annotation procedure can be out-sourced to crowdsourcing platforms like Amazon Mechanical Turk[4], but comes at a financial cost. A final reason why no one has yet undertaken these efforts for infants is that general research largely focuses on adults. The number of possible applications seems to be much higher for adults than for infants.

The above reasons may serve as an explanation why the majority of infant motion analysis approaches only scarcely evaluates the accuracy of underlying motion capture methods. We believe that the second step should not be taken before the first one, i.e., each system should first demonstrate that it is capable of accurately capturing movements before predicting outcome based on these captured movements.

[4] `https://www.mturk.com`

2.5. Requirements for an Automated Motion Analysis System

In 1997, Einspieler et al. [Ein+97] described the desired properties of a GMA as follows:

> "A quick, non-invasive, even non-intrusive and cheap method with high reliability and high validity is, therefore, most desirable and much in demand for early assessment of neurological deviations which lead to cerebral palsy and developmental deficits later on."

We believe that these requirements should hold true for automated methods, in order for them to be widely used in clinical practice. The mentioned properties can be assigned to the motion capture stage (quick, non-invasive/non-intrusive, cheap), and the motion analysis stage (high reliability, high validity), cf. Fig. 2.1. Regarding the motion capture stage, we identify the following properties for sensors:

- **Quick**: no additional effort is introduced by the recording system, like accurately placing markers or calibrating sensors before use.
- **Non-invasive/non-intrusive**: nothing that could have an influence on the infant's state of happiness/behavior/motions shall be attached to infants, like markers or sensors.
- **Cheap**: to allow widespread screening, a sensor system for hundreds of thousands of Euros is prohibitive.
- **3D**: In addition to the requirements stated above, we believe that the data is required to be three-dimensional. The real world is 3D, and 2D poses can be ambiguous. In general, reducing the number of dimensions leads to a loss of information, and we aim at measuring motion as realistically as possible.

One could argue that wireless wearables have become small enough for not being intrusive anymore. Still, even if the infant is not bothered by attached sensors, the effort of (accurately) placing the sensors on multiple predefined positions on the body remains, and has to be performed every time a recording is made. Attaching objects to an infant, especially in a clinical setting, requires these objects to be hygienic, which means they have to be cleaned/disinfected or replaced after each use. Infants may be cranky, parents have questions, examinations need to be prepared, so the time that medical staff uses to concentrate on the capture system should be kept to a minimum.

To our knowledge, the only sensor that fulfills the four properties is an RGB-D camera, which has been previously shown to be suitable for healthcare applications [Mor+16b]. Medical staff only has to press a record and stop button (quick). Infants do not notice being recorded (non-invasive). Only the sensor and a connected laptop / tablet are required. RGB-D sensors are available for few hundreds of Euros/Dollars, depending on sensor type and manufacturer (cheap). With the

Table 2.7.: Advantages and disadvantages of different sensor systems for infant motion capture. *depends on sensor choice.

Sensor	Disadvantages	Advantages
Vicon	Expensive, intrusive, setup effort, occlusion problems	High accuracy, high frame rate
Wearables	intrusive, drift issues, setup effort	low-cost, direct motion sensing, high frame rate*
RGB	2D, dependent on lighting and appearance	low-cost, easy setup, high frame rate*
2D marker tracking	2D, intrusive, marker occlusions, setup effort	low-cost, easy calculation of marker positions, high frame rate*
RGB-D	limited frame rate* (30 Hz for Kinect)	low-cost, easy setup, 3D, invariant to lighting and appearance (depth)

camera calibration (or software that often comes with a sensor), depth images can be transformed into metrically accurate 3D point clouds (3D).

An overview of advantages and disadvantages of different sensor types is presented in Tab. 2.7

In this chapter, we provided a medical description of the motor disorder we aim to automatically detect, cerebral palsy. We described clinical tools for detection of CP with a focus on the most reliable one, which is the general movement assessment. We reviewed approaches towards automation of CP detection, and in particular methods for capturing infant motion. Based on an analysis of existing approaches, we defined requirements for automated infant motion analysis systems, and concluded that RGB-D sensors seem to be the best choice regarding the requirements.

3. Recording Setup

Based on the defined requirements in the previous chapter, we selected RGB-D cameras as the sensor of choice. We now give an overview of the complete system setup and the special conditions that apply to medical infant motion analysis. We review available depth sensor types and present practical examples showing their characteristics in the infant use-case.

3.1. Setup Constraints

We defined properties of the recording setup to be used in a clinical environment to be low-cost, 3D, non-intrusive, easy to setup, uncomplicated to use and smoothly integrated in standard examination protocols. We show our recording setup that is deployed in a children's hospital in Fig. 3.1. The Kinect sensor is securely mounted above an examination table and connected to a laptop.

Several additional constraints are imposed on the recording protocol from the definition of how to perform the general movement assessment. The infants have to be in a state of active wakefulness and should be recorded in supine position. External stimulation with toys or dummies, or interaction with persons influences movements and has to be avoided. This implies that the infant can be assumed to be the only person in the image. At a young age, infants are unable to roll over, which means that we do not have to deal with back views and mostly observe frontal views. Infants should be recorded in light clothing. In our data, we observe diapers, onesies, tights, or no clothing at all.

Further constraints are dictated by common sense. The temperature should be adapted to the needs of infants, and a caregiver should be nearby. Any risk of the infant falling down from the recording area has to be eliminated.

3.2. RGB-D Sensors

3D range scanners have long been around [Bes89], but the production for consumer mass market has brought down prices so that they have the potential of becoming widespread for applications like infant motion analysis screenings. Different methods for capturing depth information are used, which we will explain in the following.

Figure 3.1.: Recording setup. RGB-D camera mounted 1 m above examination table, facing down. Connected to laptop. Scene overlaid with infant point cloud.

3.2.1. Structured Light

Structured light sensors project a pattern onto the scene (cf. Fig. 3.2 (b)), often using an infrared projector, and capture images of the illuminated scene at the same time using a camera. Any non-planar objects in the scene lead to a distortion of the pattern in the camera image. Since the pattern is known to the sensor, depth can be inferred from the distorted pattern in the image. An overview of different patterns and techniques for differentiating distorted projected light spots from images can be found in [Bes89] or [Gen11].

Probably the most famous example of a structured light RGB-D sensor is the first Microsoft Kinect sensor, in the following termed Kinect V1 (see Fig. 3.2 (a), left). It was introduced for gesture control of video games in gaming console Microsoft XBox 360 in 2010, and a Windows version was produced in 2011. An analysis of underlying components can be found in [MS13].

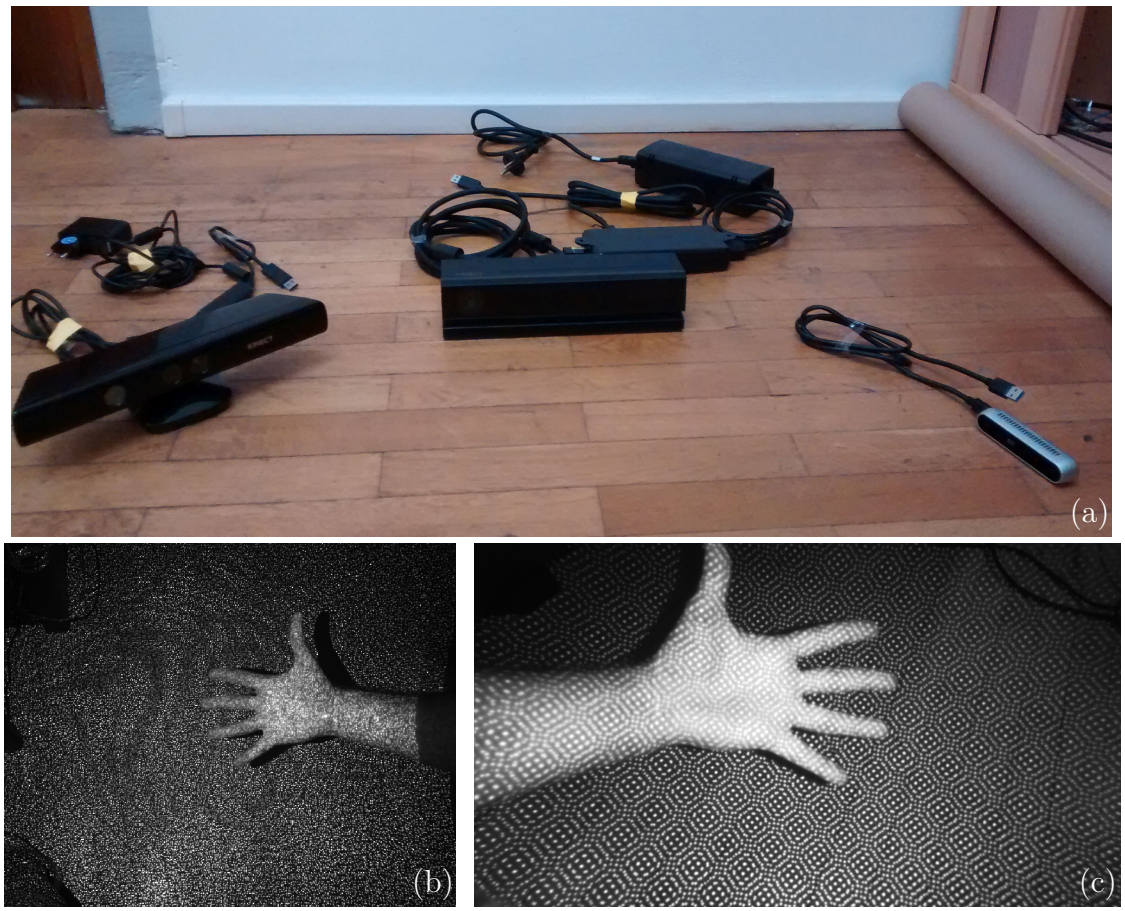


Figure 3.2.: (a) From left to right: Kinect V1, Kinect V2, RealSense D415. (b) Projected infra red light pattern of Kinect V1. (c) Pattern of RealSense D415. Note the different fields of view.

The distance between points of the projected pattern naturally increases with distance to the camera, resulting in a reduction of depth density. The depth resolution is not constant and decreases with an increase of the distance between camera and the scene [Kho11]. Since the projector and the camera – recording the pattern – can not be in the exact same place physically, there may be parts of the scene which are illuminated by the projector, but remain invisible to the camera. Pixels for which no depth values could be reconstructed are assigned a special "missing pixel" value. Missing pixels are mostly observed at object borders, but may also be caused by reflective surfaces.

3.2.2. Time-of-Flight

Time of Flight (ToF) sensors send out light flashes and measure the time it takes for the light to "bounce" back from the scene, e.g., by analyzing phase shift. Using

Table 3.1.: Depth sensor specifications. *Max. frame rate depends on resolution.

	Kinect V1	Kinect V2	RealSense D415
Max. depth resolution	640 × 480	512 × 424	1280 × 720
Depth FOV (in degrees)	58.5 × 46.6	70.6 × 60	63.4 × 40.4
Max. RGB resolution	640 × 480	1920 × 1080	1920 × 1080
RGB FOV (in degrees)	62 × 48.6	84.1 × 53.8	69.4 × 42.5
Framerate	30 Hz	30 Hz	Up to 90 Hz*
Type	Structured light	Time of Flight	Active stereo

the speed of light, the distance can be inferred. This allows for a constant depth resolution independent of distance to the camera. An independent measurement is taken for each pixel which improves the capture of fine structures compared to the structured light approach. More technical descriptions of the underlying principles can be found in [Rey+11; TB14; Lef+13].

The second generation of Microsoft Kinect, which we term Kinect V2 was released in 2013 for XBox One, and in 2014 as "Kinect for Windows" (Fig. 3.2 (a), middle). It allowed more accurate capture of free-standing people with lower levels of noise.

3.2.3. Depth from Stereo

Using two cameras with a (fixed) baseline, depth can be inferred by triangulating points that are visible in both images. By calculating and matching visual features between both views, corresponding points can be identified. However, good features for matching can only be found in parts of the scene that are textured, which is why for untextured parts of the scene either depth estimation fails or is highly unreliable.

Active stereo sensors project a texture pattern onto the scene (Fig. 3.2 (c)), but not for inferring depth from the deformation of the pattern in the camera image like structured light sensors, but for texturing the scene, so that triangulation of points on previously untextured areas becomes possible [Nis84].

This technique is applied in current Intel RealSense cameras [Kes+17]. The D400 series, including models D415 (Fig. 3.2 (a), right) and D435, has been released in early 2018.

3.2.4. Discussion

We show a summary of sensor specifications for Kinect V1, V2 and RealSense D415 in Tab. 3.1 and recorded samples of infant heads from two different viewpoints for each of the sensors in Fig. 3.3.

Based on the specifications, the D415 seems to be superior to Kinect V1 and V2, with higher depth resolution and frame rate. Together with the much smaller size and no

need for an external power source (cf. Fig. 3.2 (a)), it seemed like a big improvement over previous sensors. In experiments, however, the depth quality does not live up to the expectations raised by the specifications. Despite the high resolution, the head shape in the recorded point cloud appears bumpy (Fig. 3.3 (c)), and there is a significant amount of fluctuation between temporally subsequent frames (not shown here). Parts of the head border are "pulled" towards the background. The background table, which is flat in reality, is also recognized very bumpy (not shown), and this bumpiness increases with the distance of the camera to the recorded scene. The software accompanying the D415 is under continuous development, and many parameters exist to influence the recording properties. We have not been able to fix the issues by using different parameters and also experienced them with the RealSense D435, which seems to be better suited for longer ranges and larger scenes.

The Kinect V2 seems to have a lower depth resolution than the V1, despite being its successor. Yet, the advantage of the V2 is that for each pixel an independent measure is performed, while in the Kinect V1, the depth image is reconstructed from the structure of a known sparse pattern. The ToF technique works well for the intended gaming scenario, i.e., freely standing people (far from walls or other large objects). Nevertheless, in the infant scenario – a camera mounted above an infant lying on a flat surface – the technical property of performing distance measurements for each pixel becomes a drawback. When foreground (infant) and background (table) are close to each other, light flashes sent to the border of the infant are reflected from the infant as well as the background, leading to a distorted measurement. This phenomenon is known as "flying pixels" [Rey+11]. The distortion does not just affect a small number of pixels at the border, though, but rather seems to "suck" points lying closer to the middle of the head towards the background, as can be seen in Fig. 3.3 (b). The head shape is not round as in reality, but rather cone-shaped. It can further be noticed that some of the noisy head pixels lie behind the table plane.

The Kinect V1, although being the oldest of the three sensors, provides measurements that are visually closest to reality in the infant scenario (Fig. 3.3 (a)). A general drawback of the V1 is depth quantization due to limited depth resolution, which increases with distance of the camera to the scene. In our case, this effect is negligible, since the camera is very close to the scene (< 1m). The captured head shape is round and consistent, with small amounts of noise at the borders. Opposed to the other sensors, there are no flying pixels, and the head is clearly disconnected from the background. We come to the conclusion that, although other sensors have seemingly better specifications, the Kinect V1 currently seems the best choice for the infant use-case. However, current developments of new sensors seem very promising [Bam+18; Kow+18].

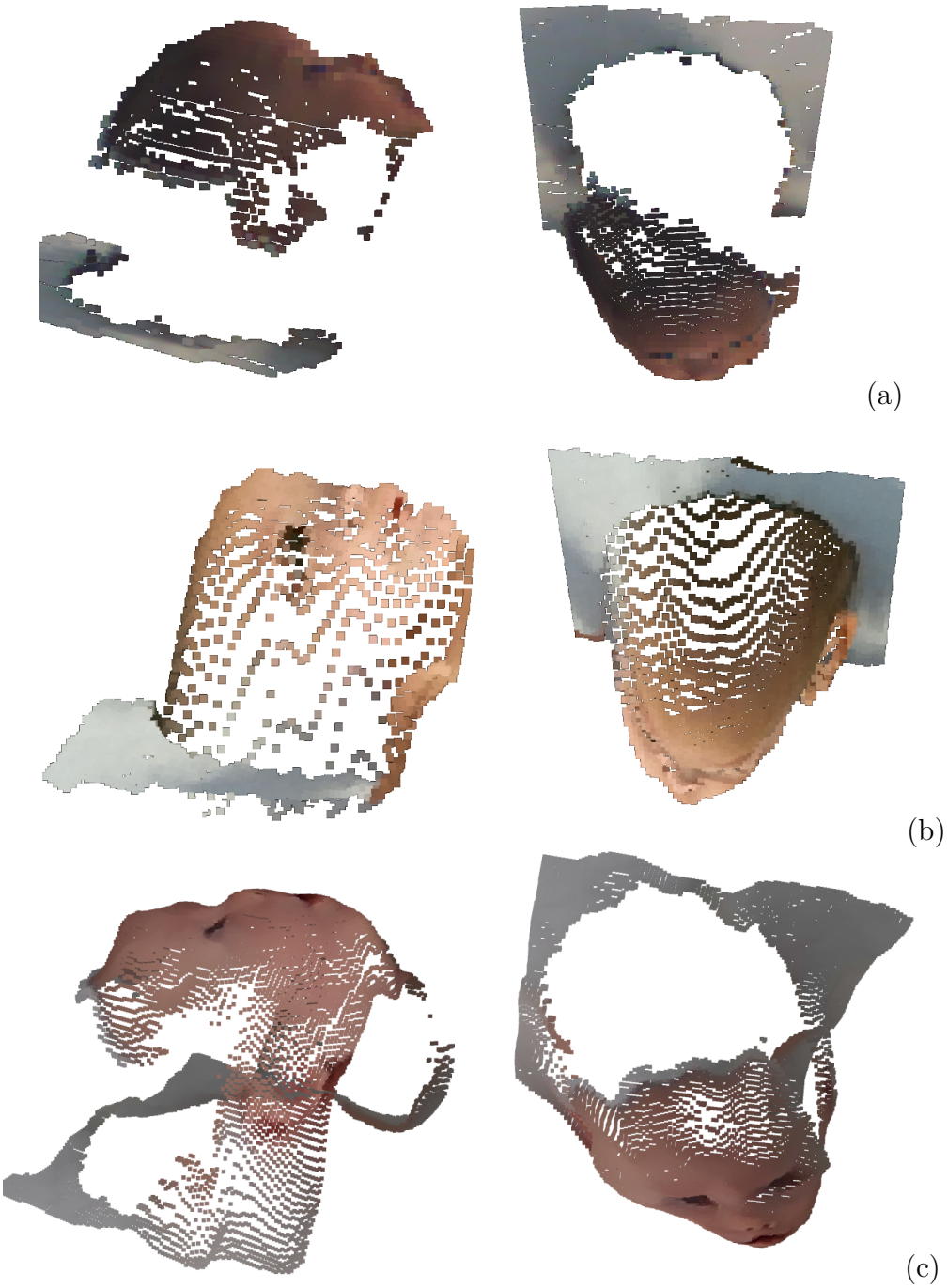


Figure 3.3.: Samples showing different infant heads. (a) Kinect V1, (b) Kinect V2, (c) RealSense D415. Left: side view, right: top view.

4. Body Pose Estimation in Depth Images using Random Ferns

Now that requirements and sensor types have been reviewed, the next step is to extract the 3D pose from RGB-D data. We review related work in the field of 3D pose estimation in depth images, and present our method for pixel-wise body part classification using random ferns. We discuss limitations of the proposed approach, and develop methods to overcome these.

Parts of this chapter were published as [Hes+15] and [Hes+17a].

4.1. Related Work – Pose Estimation from RGB-D Data

Human pose estimation in images is one of the key problems of computer vision [Sig14] and has been studied since the 1980's [LC85].

Human pose estimation from **2D RGB images** has greatly benefited from the rise of deep learning. There have been countless studies consistently advancing the state of the art of 2D pose estimation [TS14; Tom+14; Fan+15; Car+16; New+16; Pis+16; Ins+16; Wei+16; Cao+17; Ins+17; Pap+17; And+18], as well as approaches for 3D pose estimation from 2D images [Ram+12; Tek+16; Zho+16; Mar+17b; Meh+17b; Meh+17a; MN17; Pav+17; Rog+17; Sun+17; Tek+17; Tom+17; Pav+18].

To review all of them in detail is out of the scope of this work, and we will focus on the estimation of **3D pose from depth images**.

The introduction of the Microsoft Kinect sensor triggered a lot of research on pose estimation in RGB-D or depth data. Initially presented as a gesture control device for gaming console XBox, a body tracking algorithm was included. The underlying pose estimation was developed by Shotton et al. [Sho+11]. Due to hardware limitations and gaming requirements – on-board real-time processing with the ability to quickly recover from tracking failures – they decided to apply random forests to the task, which had proven their usefulness for object recognition purposes [WS06]. They chose a bottom-up approach, i.e., body parts are located first and the underlying skeleton/joint positions are then induced from the body parts. Multiple random decision trees are trained with a large number of synthetic training samples

to provide probability distributions over body parts for each input pixel. In order to quickly recover from tracking failures, this approach uses no temporal information and estimates pose in single depth images. Our approach for pose estimation in depth images has been inspired by the work of Shotton et al., which is why we describe their approach in more detail in the next section. The final Kinect sensor included an additional tracking component, which was never disclosed.

The initial success of the Kinect body tracking led to an increased use of **random forests** for pose estimation in depth images.

Buys et al. create a system for jointly detecting bodies and estimating pose, and include an appearance model to refine estimates [Buy+14]. Girshick et al. skip the intermediate body part segmentation, and use Hough forests to directly regress joint positions from depth images [Gir+11]. Jung et al. iteratively regress the direction from the center towards each joint, using a random tree walk [Jun+15]. In subsequent work, votes are cast on locating all joints using the random tree walk without identifying them, i.e., without assigning a body part label, followed by K-means clustering on the votes. Based on these clusters, a nearest exemplar retrieval is used to assign body part labels to the clusters [Jun+16].

Prior knowledge about body constraints has been used to advance the performance of random forests. Previous approaches assume independence between output variables, i.e., body parts or joint locations, which is why Sun et al. introduced conditional regression forests to model dependencies between different parts of the body and also allow the integration of prior knowledge like height or orientation of a person [Sun+12]. Xia et al. use a cascaded regression forest to model dependencies between body parts [Xia+18]. Taylor et al. apply a full body model to estimating the pose, which by design incorporates kinematic constraints [Tay+12]. The generative model-based approach is combined with a discriminative approach using random forests to predict dense correspondences between points from the depth image and the model surface. Pons-Moll et al. argue that this objective is not the "correct" one, given the task, and propose to use an objective that minimizes the uncertainty of distributions in the metric space of human body surfaces, which they call the *Metric Space Information Gain* [PM+13; PM+15b]. Lallemand et al. leverage information of the performed activity in training data to achieve a better partitioning of the pose space [Lal+13]. Park et al. rely on a two-stage random forest approach [Par+17]. Their first forest regresses joint positions from input pixels, while their second forest aims at verifying the "votes" from the first forest.

Although the use of random forests to predict the body part class or joint position from depth pixels has been the dominating approach, some researchers followed a different strategy in locating **geodesic extrema** for identifying body landmarks. Plagemann et al. detect interest points based on geodesic extrema on the body surface mesh, and train a classifier to assign body part labels to local patches extracted around interest points [Pla+10]. Baak et al. rely on a similar first step, i.e., detecting geodesic extrema [Baa+11]. These extrema are used as features to lookup

the full body pose from a pre-learned database. In addition, they perform local optimization of pose parameters of a body model initialized with the pose from the previous frame. They fuse these two hypotheses using a voting scheme to produce a final pose estimate. Geodesic distances are also used by Schwarz et al. [Sch+11], but they reinforce their method by applying optical flow to distinguish self-occluding body parts.

A different approach is to transform 3D point clouds into a **voxel representation** (3D pixel) before processing. Schick and Stiefelhagen apply pictorial structures to the task of 3D pose estimation and use a supervoxel segmentation to make the method tractable in 3D [SS15].

More recently, **convolutional neural networks** (CNNs) have been applied to pose estimation from (voxelized) depth images. Instead of using the complete depth map as input, Haque et al. generate "glimpses" of the depth image, which are images where a small window is displayed at full resolution, but the resolution decreases with distance to the center [Haq+16]. This is supposed to make their model focus on specific parts while keeping general spatial information. They convert each glimpse into a 3D voxel grid, and learn an embedding into a viewpoint-invariant feature space. Based on these features, they train a multi-task network to determine the visibility of a body part and to predict body joint locations. Zimmerman et al. predict 3D pose and hand normal vectors [Zim+18]. They extract 2D keypoints from RGB images and generate a voxelized occupancy grid from the corresponding depth image. The 2D estimates are then "lifted" to 3D using a CNN. Moon et al. argue that mappings from 2D depth to 3D coordinates are highly non-linear and complicate the task of pose estimation [Moo+18]. For this reason, they propose a voxel-to-voxel CNN, which takes a 3D voxelized grid as input and estimates the per-voxel likelihood for each body joint. Marin et al. base their approach on the assumption that every pose can be represented by a linear combination of a number of prototype 3D poses [MJ+18]. They train a CNN to predict the weights for combining multiple prototypes to form the final 3D pose estimate.

In literature, the random forest approach has prevailed for pose estimation from depth images. Good performance, together with simplicity and efficiency has inspired the development of many variants.

4.2. Random Ferns

As described in the previous section, the body tracking of the Kinect V1 has triggered a whole body of work based on random forests. The initial approach for the Kinect was introduced by Shotton et al. [Sho+11]. We briefly recap this method because the underlying ideas are similar to the approach we present in this section. In general, the depth image is assumed to be segmented to only contain pixels corresponding to a human body.

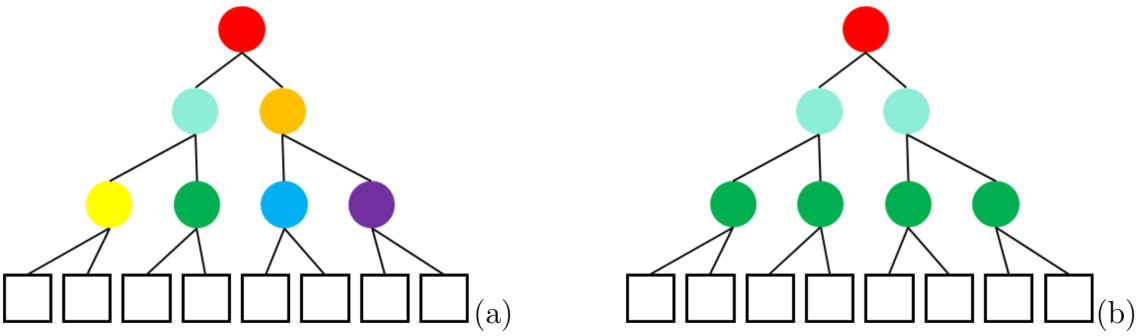


Figure 4.1.: Structures of different kinds of decision trees. Colored dots depict split nodes containing features, squares depict leaf nodes containing probability distributions. (a) Standard binary decision tree. The decision which node will be traversed next depends on the outcome of the current node. (b) Fern. Always the same nodes are traversed (indicated by colors), irrespective of the outcome of individual nodes.

4.2.1. The Pose Estimation Method Inside the Microsoft Kinect

Instead of directly estimating body pose in depth images, an intermediate representation is added, i.e., they assign a body part label to each pixel of the depth image. From these body parts and the known joint positions, a mean shift mode finding method is applied and a z offset is learned to push joints from the body surface inside the body.

They chose a *random forest* approach for the task of pixel-wise body part labeling, which has been successfully applied to computer vision tasks like object recognition [WS06] and can be very efficiently implemented to allow real-time performance [Sha08]. The basic idea is to merge many weak learners into one strong learner. A random forest consists of multiple decision trees. Each tree consists of split nodes and leaf nodes. In each split node, the input is evaluated with respect to a feature and a threshold, and depending on the result of this evaluation, either the left or right branch is taken (in a binary tree). The tree is traversed until a leaf node is reached, in which a probability distribution over output classes is stored. A sample decision tree is depicted in Fig. 4.1 (a).

During training, a set of candidate features and thresholds is randomly sampled for each node – hence the name *random* forest. This is done because a brute force approach of testing all possible features is computationally prohibitive. The feature candidates are evaluated w.r.t. an objective function, taking into account the balance of left and right partitions. The best candidate is selected, and the process is repeated until a predefined level of tree depth is reached. The number of candidates to be evaluated grows exponentially with tree depth, which makes the training procedure computationally expensive for deep trees.

In the approach by Shotton et al., the input to a tree is a depth pixel, the features

are binary depth comparisons in the neighborhood of the input pixel, and the leaf nodes store probabilities over body part classes. The objective function to evaluate candidate features during training is the Shannon entropy. To train the classifier, a large amount of training data is needed, which the authors generate by applying many poses acquired by marker-based motion capture recordings to a 3D body surface model. They divide the body model into 31 parts and render labeled depth images from the posed model. In the leaf nodes, probability distributions over these body part classes are accumulated from the labeled input pixels reaching the respective leaf nodes during training. This process is repeated for a defined number of trees, which form the forest. At test time, each input pixel traverses the trees, and is labeled with the combined probability distributions of the reached leaf nodes of all trees.

Three trees of depth 20 are trained using one million images and a random subset of 2000 pixels per training image. The training takes one day on a 1000 core cluster.

The body tracking that was released in the Kinect SDK is based on this approach, and has been used for many applications like gait analysis [Sun+14] or Alzheimer's disease assessment [Iar+14]. However, the system is not applicable for the purpose of tracking infants since the minimum size of tracked humans is one meter. The system was not trained on smaller persons as the main application is the gaming scenario.

4.2.2. Pose Estimation from Pixel-wise Body Part Predictions

The random forest approach has shown to work well on challenging data under real-time constraints. Therefore, re-training with infant data seemed like a promising approach. However, 1000 core clusters are not widely available, or expensive to rent. The parallelized implementation poses many challenges, which are described by Budiu et al. [Bud+11]. Ignoring the processing overhead, running the training on a single core would take 1000 days, which is infeasible.

Therefore, our goal is to reduce the training time while keeping pose estimation accuracy. We chose to rely on a random ferns body part classifier, which has been shown to be an efficient and robust alternative to random forests [Ozu+07]. Similar to [Sho+11], we estimate body part labels for each pixel of the input depth image.

Ferns are a simplified version of decision trees. The biggest difference is that ferns evaluate the *same* feature in *all* split nodes of a tree level (cf. Fig. 4.1). In decision trees, the evaluation of the candidate set for the next feature depends on the output of all previous nodes in the tree leading to the current node. In ferns, the order of evaluation is irrelevant, since always the same features are evaluated, irrespective of the traversed path. The usage of ferns greatly decreases training time. However, the reduced complexity of ferns naturally leads to a degradation of performance. To compensate for that, we train many ferns instead of few trees.

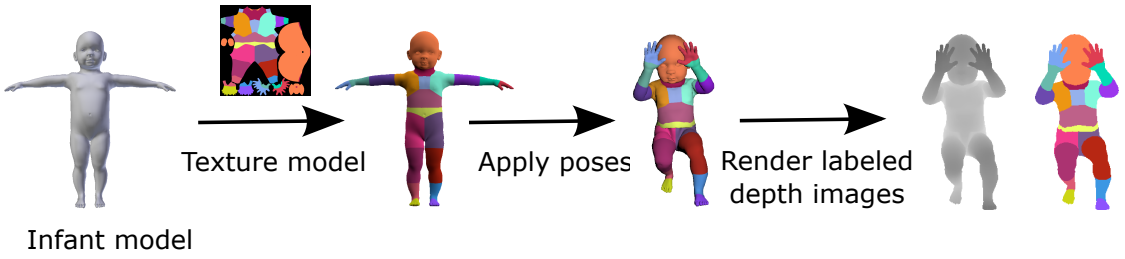


Figure 4.2.: Generation of synthetic training data. Infant model is textured with colors corresponding to body parts. After different poses are applied, labeled depth images are generated that serve as input for the fern training.

Binary Features. The task at hand is to correctly estimate the body part class of each pixel of a depth image displaying a human. We do this by applying depth comparisons between the current input pixel and several pixels within a predefined neighborhood radius. We define each of these comparisons with an associated threshold to be a feature. The outcome of the feature is either 1 or 0, depending on the result of the depth comparison being greater than the threshold or not.

We use depth comparisons similar to [Sho+11], defined as

$$z_\phi(I, x) = d_I(x) - d_I(x + \phi \cdot r(x)), \tag{4.1}$$

where ϕ is the relative offset to a given pixel x in image I, $d_I(x)$ returns the depth of x and $r(x) = \frac{foc}{d_I(x)}$ is a normalization factor for depth invariance, where foc is the focal length of the depth sensor in pixels. A binary feature consists of a depth comparison in combination with a threshold τ, and is evaluated as

$$f_\phi(I, x) = \begin{cases} 1, & \text{if } z_\phi(I, x) > \tau, \\ 0, & \text{else.} \end{cases} \tag{4.2}$$

Each binary feature for itself is not very meaningful, therefore we combine many features to obtain an accurate estimate.

Ferns. We follow the formulation of [Ozu+07].

Let $c_i, i = 1, \ldots, H$ be the set of classes and let $f_j, j = 1, \ldots, N$ be the binary features. We want to find

$$\hat{c} = \arg\max_{c_i} P(C = c_i | f_1, f_2, \ldots, f_N), \tag{4.3}$$

where C is a random variable representing the body part class. Applying Bayes' formula leads to

$$P(C = c_i | f_1, f_2, \ldots, f_N) = \frac{P(f_1, f_2, \ldots, f_N | C = c_i) P(C = c_i)}{P(f_1, f_2, \ldots, f_N)}. \tag{4.4}$$

In contrast to [Ozu+07], who assume a uniform prior $P(C)$, we use prior probabilities depending on the number of pixels representing each body part in the training data. Being independent of the class, the denominator is regarded as a scaling factor and therefore omitted.

A complete representation of the joint probability of all features requires storing and estimating 2^N entries for each class, which makes it computationally intractable for more than a few classes. If we assume complete independence of the features, the problem becomes trivial, but the correlation between features is ignored. Therefore, a compromise is chosen by partitioning the set of all features into M groups of size $S = \frac{N}{M}$, where each group is called *fern*. In each fern, the joint probability is computed. This leads to

$$P(f_1, f_2, \ldots, f_N | C = c_i) = \prod_{k=1}^{M} P(F_m | C = c_i), \tag{4.5}$$

where $F_m = \{f_{\sigma(m,1)}, f_{\sigma(m,2)}, \ldots, f_{\sigma(m,S)}\}, m = 1, \ldots, M$ represents the m^{th} fern and $\sigma(m, j)$ is a random permutation function with range $1, \ldots, N$.

This semi-naive Bayesian approach models only some of the dependencies between features, but can be handled easily, as we need to store and estimate $M \times 2^S$ parameters. In all our experiments we use $M = 15$, $S = 12$. Each fern can be seen as a special kind of decision tree, with binary features as split nodes, where within each level of the tree the same features are evaluated. The depth of the tree corresponds to the group size S. In each leaf node, the probability distribution of all classes is stored. The complete set of M ferns is called an *ensemble of ferns*.

Training data generation. We synthetically create a large amount of labeled data for training the ferns. We use a 3D body model of an infant from MakeHuman [Mak], an open source tool for making 3D characters. A subset of the available body joints is selected in our experiments. The joints do not necessarily correspond to real joints of the human body, but serve as a tool for dividing the model into different regions. We choose 21 joints for infants: head, neck, shoulders, elbows, hands, fingers, upper body center, body center, stomach, hips, knees, feet and toes. A texture is applied to the model that maps each skin pixel to a color according to the closest joint. The open source software Blender [Ble15] is used for animating the model in many different poses, using the CMU motion capture dataset (CMU MoCap) [GS01]. We rely on this data set because no data set containing infant motion is available. The dataset contains a variety of poses that were captured by a Vicon system at 120 Hz. Consecutive poses are very similar due to the high capture rate and are omitted if the summed joint distances lie below a predefined threshold. We generate depth images from the posed model and assign a body part label to each pixel according to the model texture. The virtual camera viewpoint from which the depth images are generated can be chosen arbitrarily. We use mainly frontal views of the body as we assume the infants to be lying in supine position. The data generation procedure is illustrated in Fig. 4.2.

Fern Training. The algorithm for the training procedure is outlined in Algorithm 1. The goal is to build an ensemble E consisting of M ferns. We start by creating a fern, given its depth S and the neighborhood radius for pixel offsets. The randomness is introduced in the sampling of the binary features, i.e., pixel offsets within the specified neighborhood and thresholds (as used in Eq. 4.1 and 4.2). The outcome of the features in the fern accumulates to the descriptor $F_m = (f_{\phi 1}, f_{\phi 2}, \ldots, f_{\phi S})$ and is indexed by a binary code that indicates which leaf node is reached by the input pixel. In each leaf node, the probability distribution over all body part classes is stored and is given by

$$P(F_m = k|C = c_i) = \frac{n_{k,i} + u}{\sum_k (n_{k,i} + u)}, \tag{4.6}$$

where we consider F_m to be equal to k if the binary code of the feature descriptor equals k. Furthermore, $n_{k,i}$ is the entry in the histogram of descriptor F_m and is equal to the number of pixels belonging to class i that evaluate to fern value k. The variable u can be seen as a Dirichlet prior to avoid probabilities of zero, in case a leaf is not reached by any input pixel, which makes the multiplicative combination of all ferns zero. Choosing $u = 1$ leads to

$$P(F_m = k|C = c_i) = \frac{n_{k,i} + 1}{N_i + K}, \tag{4.7}$$

with N_i being the total number of pixels in the training data that belong to class i.

We evaluate the error that results from classifying all input pixels using the ferns in the ensemble together with the current fern: as before, every pixel of every input image is input to each of the ferns and the resulting probability distributions of all ferns are multiplied. From the resulting distribution, the class with highest probability is the estimated class for the current input pixel. The estimate is compared to the ground truth and the average error over all pixels in all images is computed.

The described steps are repeated i_F times, after which the fern generating the lowest error rate together with E is added to E. After M ferns are added to the ensemble E, the algorithm terminates.

From body parts to joint positions. After passing the ensemble of ferns, each pixel of the body is assigned to a body part class. We filter out assumingly incorrectly labeled pixels by creating connected clusters in the depth image from equally labeled pixels and ignoring all pixels not belonging to the largest cluster of their respective body part class. We infer the joint positions by calculating the mean of the remaining cluster of each body part. A drawback of this method is that the joints lie close to the body surface, which does not resemble a real human skeleton. The incorporation of a procedure for locating joint positions in a way that they conform to a human body with respect to bone sizes and locations would probably improve accuracy.

Algorithm 1 Fern training procedure

Input: M: predefined size of fern ensemble
i_F: number of iterations per fern
S: depth of fern
R: size of neighborhood radius
I: labeled depth images
Output: ensemble of ferns E

Initialize:
2-D Array $Histogram$ of size (#leaf nodes (2^S)) $\times$ (#classes) with all zeros
F_{best} = NULL
$Err_{min} = \infty$
for $i = 0$ to M **do**
  for $j = 0$ to i_F **do**
    $F :=$ CreateRandomFern(S, R)
    for all images $im \in I$ **do**
      for all pixels p in im **do**
        $k :=$ GetLeafNodeIndex(F, p)
        Set $Histogram[k]$[GetLabel(p)] += 1
      end for
    end for
    $Err :=$ EvaluateTrainingError($E \bigcup F$)
    if $Err < Err_{min}$ **then**
      $F_{best} = F$
      $Err_{min} = Err$
    end if
    j += 1
  end for
  $E = E \bigcup F_{best}$
  i += 1
end for

Table 4.1.: Average error in mm per joint/body part over all sequences of PDT13 dataset.

Joint/body part	Head	Neck	Stomach	HipC	HipL	HipR
Avg. error	96	70	77	80	95	101
Joint/body part	KneeL	KneeR	FootL	FootR	ToesL	ToesR
Avg. error	114	116	113	125	154	172
Joint/body part	ShoulderL	ShoulderR	ElbowL	ElbowR	HandL	HandR
Avg. error	97	90	166	146	248	332

Table 4.2.: Average joint positions error in mm per sequence for PDT13 dataset. The column containing M1-3 and F1-2 specifies the subject, the row containing D1-D4 the sequence.

	Kinect SDK				Our approach			
	D1	D2	D3	D4	D1	D2	D3	D4
M1	59	86	138	160	92	129	153	228
M2	38	67	81	183	70	119	117	220
M3	43	85	87	133	69	134	124	154
F1	54	98	112	187	88	142	178	178
F2	36	58	79	130	53	99	105	154
Mean	96				130			

4.2.3. Evaluation

We evaluate our method on a public pose estimation data set containing data from adults in order to compare it to the Kinect SDK. To show the usefulness of the system for the proposed application we also evaluate it on manually annotated recordings of an infant consisting of more than 1000 depth images. As a measure of accuracy we use the average distance between estimated joints and ground truth in all experiments. We further evaluate the method on our newly created MINI-RGBD data set (chapter 6). An overview of data sets that are used for evaluation can be found in Tab. A.1. We experimentally determined the parameters for training the ferns to find a good tradeoff between speed and accuracy.

Results - PDT13 dataset. We evaluate our system on the publicly available Personalized Depth Tracker Dataset (PDT13) [Hel+13a]. It offers Kinect depth recordings of 5 different adults, each performing 4 movement sequences of increasing difficulty. Ground truth joint positions were generated based on measurements of a full-body laser scanner.

We have trained our system using 735 K labeled depth images. We set the number of ferns M to 15, the depth of each fern S to 12, the number of iterations per fern i_F to 32, and the neighborhood radius for pixel offsets R to 80 cm. The complete training

Table 4.3.: Average error in mm per joint/body part in infant recording FernSeq1. Average over all joints: 41 mm.

Joint/body part	Head	Neck	Body center	HipL	HipR
Avg. error	37	20	30	12	33
Joint/body part	KneeL	KneeR	FootL	FootR	ShoulderL
Avg. error	49	45	30	28	73
Joint/body part	ShoulderR	ElbowL	ElbowR	HandL	HandR
Avg. error	27	20	24	149	44

takes about 9 hours on a 32 core workstation and the body part classification runs in real-time on a desktop PC.

As the underlying skeleton differs from ours, we apply a calibration step to remove the constant offset from our estimation to the ground truth. We take our estimate of an "easy" pose (e.g., T-pose, with extended arms and legs) once for each subject and calculate the offset to the ground truth. We add that offset to all our estimates for all sequences of that subject.

Tab. 4.2 shows a comparison with the Kinect SDK, indicating the average joint position error (AJPE) per sequence. We compare our results with those of the Kinect SDK, as both approaches are constructed in a similar manner opposed to the approach of [Hel+13a], who use a more complex model fitting process. The reader is referred to the original work for their results. Our approach performs slightly worse than the Kinect SDK, and the average joint error over all sequences is higher than that of the Kinect SDK, with 130 mm compared to 96 mm. However, it should be noted that the Kinect SDK uses an undisclosed tracking component, whereas our approach does not leverage temporal information.

In Tab. 4.1 the average error over all sequences is given for each joint/body part. Our system lacks accuracy when it is confronted with challenging limb poses and body rotations leading to occlusions, which results in the increased distance error.

Results - FernSeq1 infant recording. Infants at the age of 3 months are sized around 60 cm in average. We generate 180 K synthetic depth images using a body model of appropriate size as described in Sec. 4.2.2. We train an ensemble of ferns containing 15 ferns of depth 13 and set the pixel offset neighborhood radius to 20 cm.

We annotated a Kinect V2 recording of an infant of size 60 cm containing 1082 frames. The background of the depth image is removed prior to the pose estimation, so that it only contains the infant. It is lying on the back at a distance of 90 cm to the camera. Fig. 4.3 (a) displays the annotations to illustrate the bodily movements observed in the recording. While the body center remains in a steady position, movements of both arms and legs are observed. The infant in the recording is

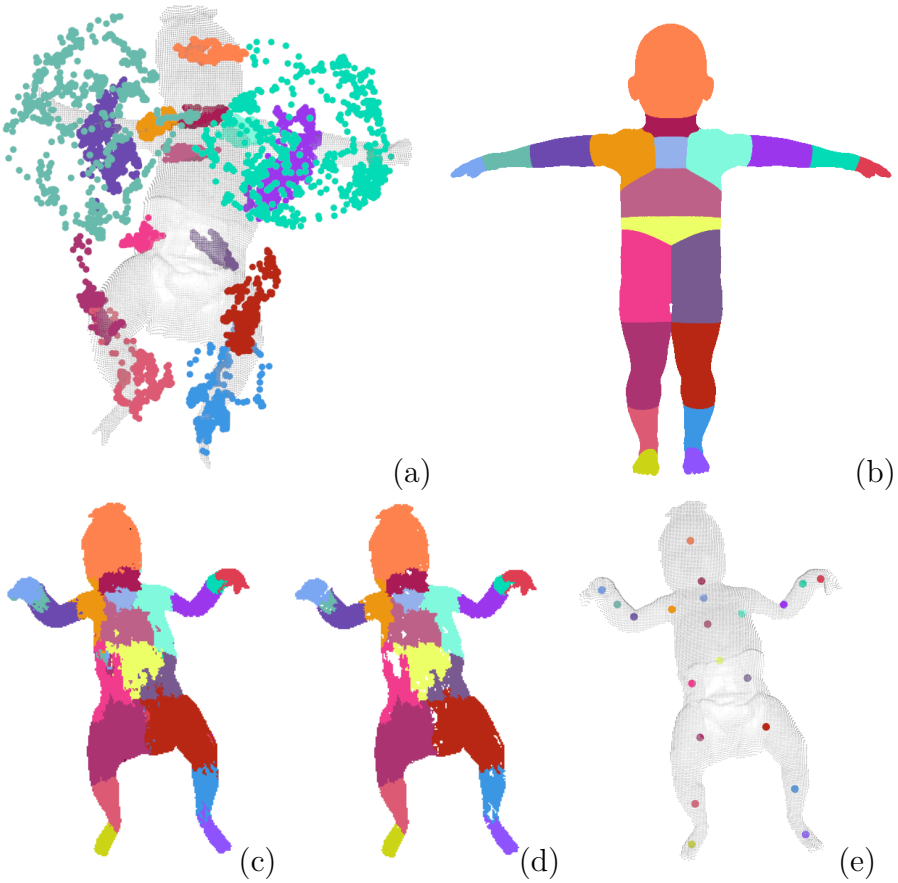


Figure 4.3.: (a) Annotations of joints illustrating infant movements in sequence FernSeq1. (b) Ground truth body part labels. (c) Body part classification sample on infant data. Estimated left shoulder and knee regions are too big, possibly due to the diaper not being present in the training data. (d) Filtering removes wrong classifications. (e) Estimated joint positions.

wearing a diaper, which makes the hips and thighs look much wider than those of the ground truth model.

Fig. 4.3 depicts an illustrative sample of the pose estimation. We show the ground truth body part labels, the estimate and the filtered estimate of body part labels in Fig. 4.3 (b), (c) and (d). It can be observed that the body regions are found in the correct positions and the filtering removes wrong classifications. The size of some regions differs from the ground truth, e.g., the left shoulder, the hips and the knees. The estimated joint positions are displayed in Fig. 4.3 (e).

Tab. 4.3 lists the average distance error per joint. The left hand shows the largest average error of 149 mm, followed by the left shoulder with 73 mm. Fig. 4.5 displays the joint distance error per frame for all joints. The error for the left hand jumps between up to 40 cm and less than 5 cm distance to ground truth.

There are two main reasons for the failure cases of the left hand. If the infant

(a) (b)

Figure 4.4.: Sample of body part estimation with high error for hand from FernSeq1. (a) Before filtering. (b) After filtering. Filtering removes correctly labeled hand as there is a bigger region at the knee with the same label. Left hand is colored in light green, left fingers in red.

pulls up the knee sideways, the filtering sometimes removes the correctly detected hand because there is a bigger region labeled 'hand' on the knee (see Fig. 4.4). This problem occurs in frames 0 to 100 and frames 830 to 1082. As soon as the size of the wrongly detected body part exceeds the size of the correctly detected part, the joint is positioned at the wrong location, resulting in a jump of the distance error. The training data does not contain such poses which may be a reason for the misclassification. The second problem we encounter occurs if the infant puts the hand very close to the body, which is the case in frames 190 to 350. The hand merges with the body, and the predicted hand is positioned on the knee. The system shows a steady and accurate performance with the distance error of the vast majority of joints hardly exceeding 5 cm. The overall average joint position error is 41 mm.

Results - MINI-RGBD data set. We evaluate our approach on the synthetic data set we created, and which is detailed in chapter 6. It consists of 12 sequences of moving infants, each containing 1000 RGB and depth frames, with ground truth 3D joint positions. The sequences are numbered in their order of difficulty, with lower numbers assigned to sequences that are considered easier.

In addition to the previously used average joint position error (AJPE), we want to evaluate overall correctness of pose estimation. We use the PCKh metric [And+14], which is commonly used for the evaluation of pose estimation approaches [Cao+17; Wei+16; Haq+16; Sci+17]. PCKh stands for the percentage of correctly identified keypoints w.r.t. the head segment length. An estimated keypoint is considered correct if its distance to the ground truth is less than 50% of the head segment length (PCKh), e.g., the distance between neck joint and head joint. The model with which the MINI-RGBD data set was created has a very short head segment

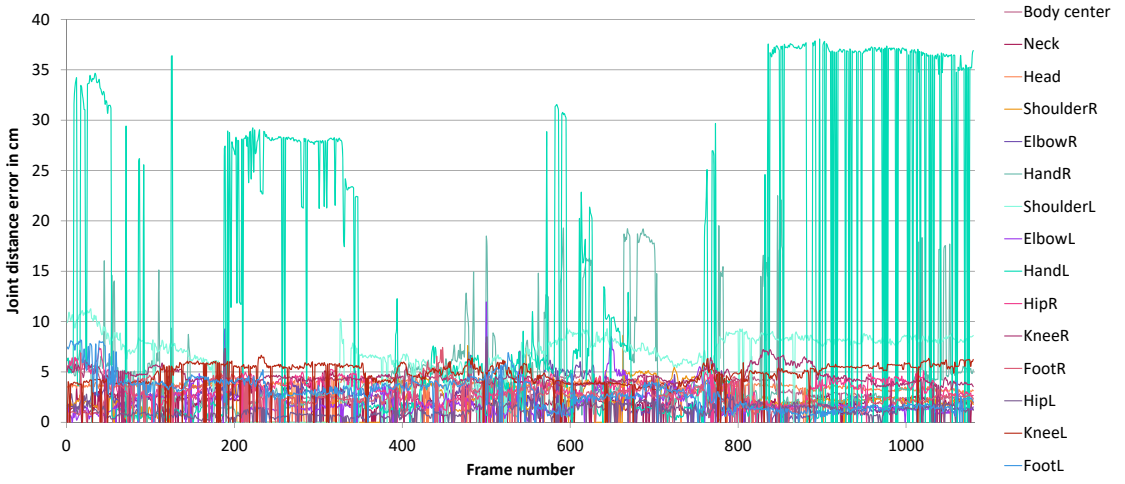


Figure 4.5.: Joint distance error per frame for all joints in infant sequence FernSeq1.

(head joint to neck joint, cf. Fig. 6.2, (c) and (f)), which is why we present results using the full head segment length (PCKh 1.0), as well as two times the head segment length (PCKh 2.0) as thresholds. Average 3D head segment length over all sequences is 2.64 cm. The head segment length for each sequence is calculated from the ground truth joint positions in an additional T-pose frame, containing the shaped model in a pose with extended arms and legs. We calculate the PCKh values for each joint for each sequence, and average numbers over all sequences, respectively over all joints. To account for differences in skeletons between our approach and the model, we calculate joint offsets for neck, shoulders, hips, and knees from the T-pose frame (Sec. 6.3), and add these offsets to the estimated joint positions in every frame.

We present results for the PCKh measure in Fig. 4.6 and Fig. 4.7, and AJPE in Fig. 4.8 and Fig. 4.9.

It can be observed that the torso joints are located with much higher accuracy than limbs. The PCKh results show that most torso classes achieve well over 80% of PCKh 2.0, with spine, chest, nect and head even approaching 100% (Fig. 4.6). The more strict PCKh 1.0 measure leads to a degradation mostly between 15 and 20% compared to PCKh 2.0. A similar trend can be observed for the evaluation per sequence (Fig. 4.7).

The discrepancy is especially obvious in the AJPE per joint (Fig. 4.8), with an AJPE of 2 to 3 cm for torso joints, and between 4 and 10 cm for limb joints. The average AJPE over all joints and sequences is 4.66 cm. Best results per sequence are obtained on sequence 2, with a PCKh 2.0 of close to 100% and an AJPE of less than 2 cm, and worst results on sequence 12, with a PCKh 2.0 of just 40% and an AJPE of nearly 10 cm (Fig. 4.7 and Fig. 4.9).

In general, we observe good results for easy, frontal poses, but with more challenging

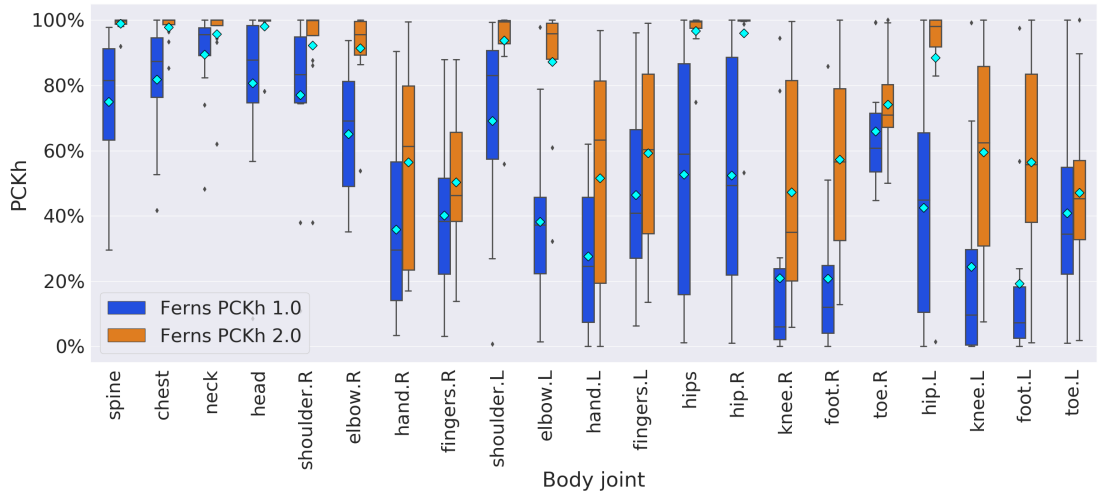


Figure 4.6.: Results for 3D pose estimation based on random ferns. Percentage of correct keypoints in relation to head segment length (PCKh) per joint as boxplot. Cyan diamonds depict the mean value.

poses including self occlusions, performance severely degrades. Next, we examine the reasons for these limitations.

Due to the lack of infant motion capture data sets, we used poses from the CMU MoCap data set, which contains adults performing a variety of everyday actions. However, the repertoire of poses shown by lying infants differs widely from those in the training set, and the method does not seem to generalize well to unseen poses. Additionally, when hands move close to the body, the estimates for the hand region pixels change to torso classes.

The approach is not rotation invariant, i.e., if the infant rotates around the vertical body axis in the image, and with that rotates out of the range of training samples, the classification performance degrades.

Kinematic constraints are not enforced, which is why impossible body part configurations (e.g., hand cluster at knee, without connection to other arm clusters) are not penalized.

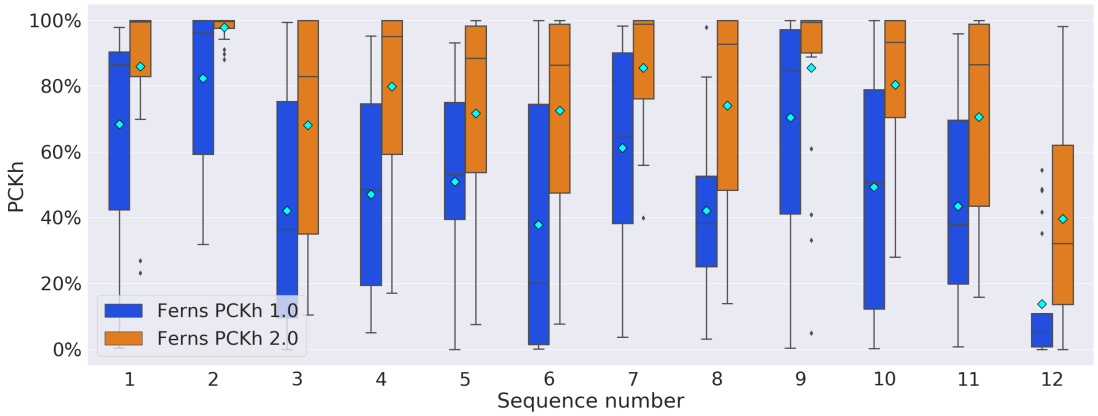


Figure 4.7.: Results for 3D pose estimation based on random ferns. PCKh per sequence. Cyan diamonds depict the mean value.

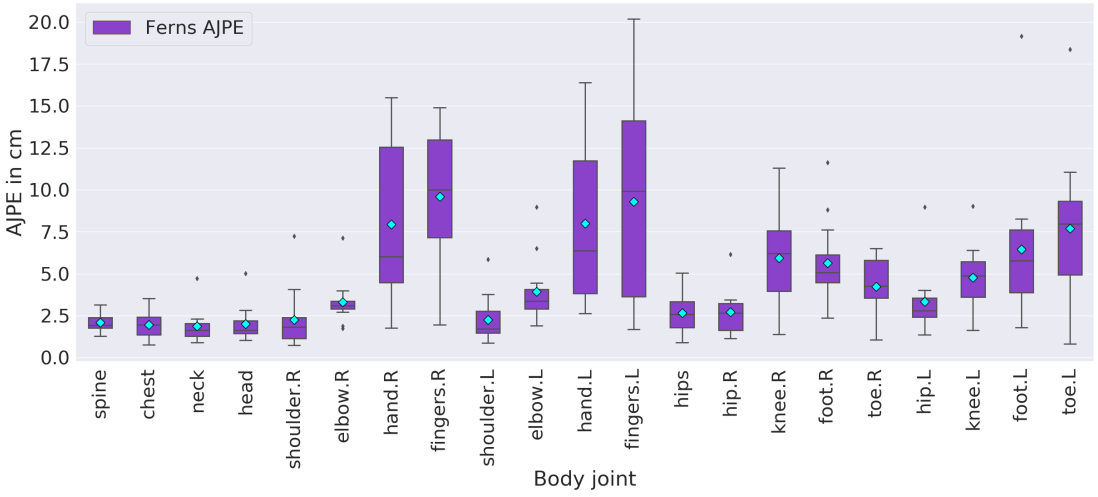


Figure 4.8.: Results for 3D pose estimation based on random ferns. Average joint position error (AJPE) per joint. Cyan diamonds depict the mean value.

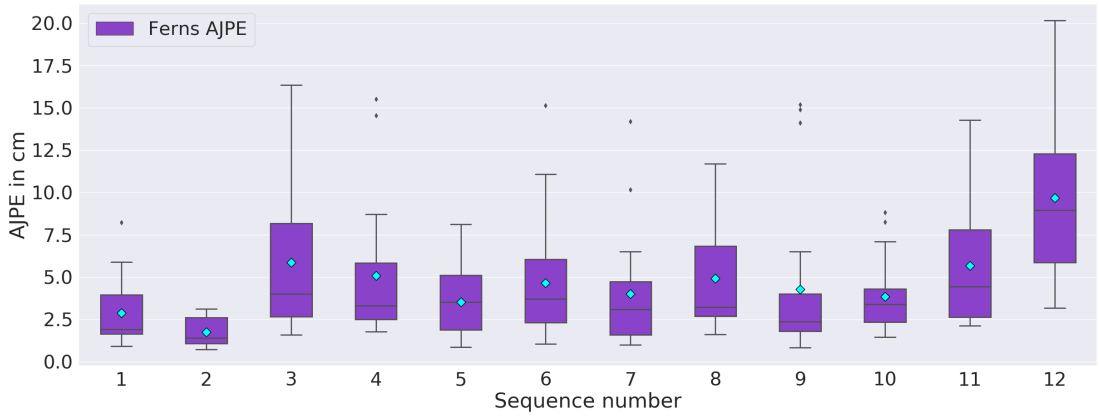


Figure 4.9.: Results for 3D pose estimation based on random ferns. Average joint position error (AJPE) per sequence. Cyan diamonds depict the mean value.

4.3. Random Fern Extensions

We propose multiple extensions to the random ferns approach, which are motivated by the previously identified limitations.

4.3.1. Multi-view Ferns

Infants at a suitable age for GMA generally are not able to turn. In order to use the approach for older children or adults who are able to walk and turn around, we examine the use of view-dependent ensembles of ferns.

We train three ensembles of ferns using data from different views: front (-45 to +45 degrees body rotation), left (-135 to -45 degrees) and right (45 to 135 degrees). The training data is produced to simulate an older child, using a human body model of size 130 cm.

We experimentally determined the following parameters: the depth of the ferns is 12, we use 15 ferns per ensemble and run 64 iterations per fern. The neighborhood radius is set to 40 cm. Each fern is trained with 60K synthetic depth images.

For evaluation, we manually annotated a sparse subset of frames from a short sequence *FernMultiView* of a child walking from left to right and back. The recording is manually divided into different sections (frontal, left, right, back) according to the side the child turns towards the camera.

Fig. 4.10 shows how the specialized ferns outperform the others in frames in which the child is seen from the corresponding view. Not every frame of the sequence is annotated, which is why the front/back view is skipped sometimes when turning from left to right (and back). For each section, the respective specialized fern ensemble shows an average joint distance error of 5 to 10 cm, compared to distances of 15 to 25 cm for the others. These preliminary results support the assumption that many specialized classifiers can improve the overall accuracy over one very general classifier. Our system allows fast training and therefore can generate different classifiers quickly to suit varying requirements. We plan to further explore the specialization of fern ensembles to specific tasks.

4.3.2. Training Data Generation

Evaluations in Sec. 4.2.3 showed failure cases for incorrectly classified body parts, e.g., knee pixels classified as hand. We show how more suitable training data can lead to an enhancement of classification quality.

The poses used in the training from the previous section are taken from the CMU MoCap database, which contains adults performing everyday activities. They are mapped to a synthetic body model of an infant. However, these poses are not

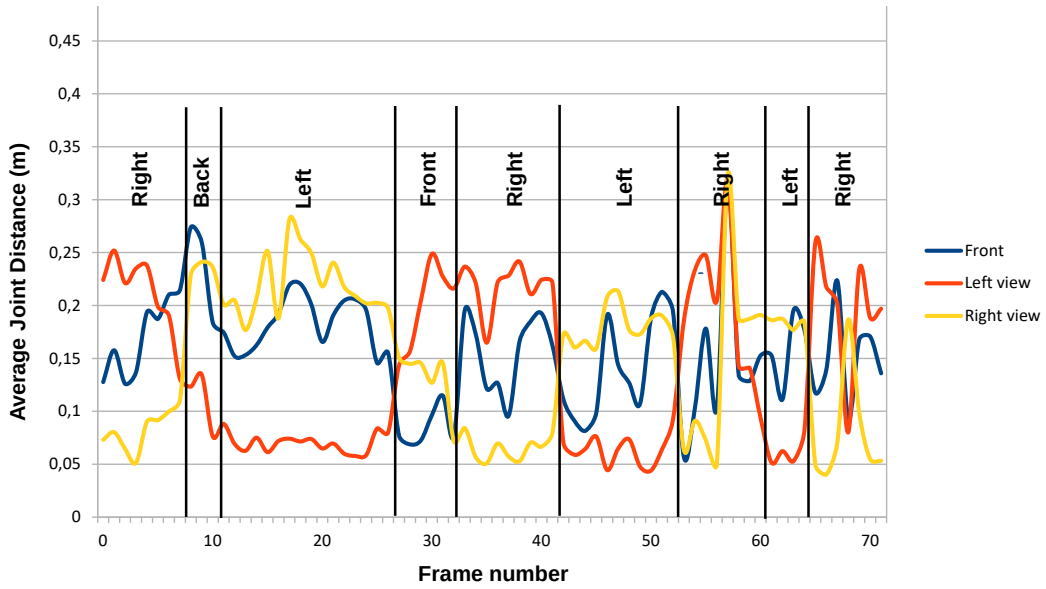


Figure 4.10.: Average joint distance error for three ferns, trained for different viewpoints, for sequence FernMultiView. Sections Front, Left, Right, Back were manually annotated for a subset of frames of a longer recording, which is why the labels change directly from left to right and back.

typical for infants and their specific motions. This bias in training data will severly decrease the classification accuracy of the ferns. As it is unfeasible to record a large amount of motion capture data from infants performing different poses, we have synthetically generated a wide range of what we consider baby-like poses. We determine angle ranges for the extremities which seem to represent infant poses realistically by visual inspection. We generate 30K poses by randomly combining angles within these ranges. We pick small random angles between 10 and -10 degrees for joints that do not belong to extremities. We generate labeled depth images from three different frontal viewpoints for each of the poses. To ensure that the training data is not biased towards either side, we mirror the generated data, resulting in an overall number of 180K training images.

4.3.3. Feature Selection

The inspection of trained ferns reveals that binary depth comparison features are included in the ferns for which pixel offsets ϕ or depth thresholds τ take values that lead to an evaluation of all training pixels to exclusively one side (all 0 or all 1). Leaf nodes which are on the wrong side of that particular feature will never be traversed during training, and therefore do not contribute to classification. Simply removing features would reduce memory requirements, but to improve classification, replacement by new (better) features is necessary. Replacing features in an existing

fern enforces recalculation of probability distributions in all leaf nodes, due to the property of ferns that each feature influences all leaf nodes. Therefore, our choice is to filter out redundant features *before* training to keep training times low and to avoid re-processing. We randomly generate a large set of features and evaluate them on the training data once using the information gain measure (Fig. 4.11). Only features with information gain above a user-specified threshold are added to a set from which features are drawn randomly during training. The information gain measure is often used for evaluating candidate features in training random decision trees (e.g., [Sho+11]). It is defined by

$$ig(\theta) = \sum_{d \in \{L,R\}} \frac{|S^d(\theta)|}{|S|} I(S^d(\theta)), \tag{4.8}$$

where $\theta = (\phi, \tau)$ represents the feature parameters, with pixel offsets $\phi = (\phi_x, \phi_y)$ and depth threshold τ. S is the set of all training pixels, which is divided into two subsets S^L and S^R according to the split induced by θ. I is the Shannon entropy of the distribution of body part classes corresponding to pixels in S:

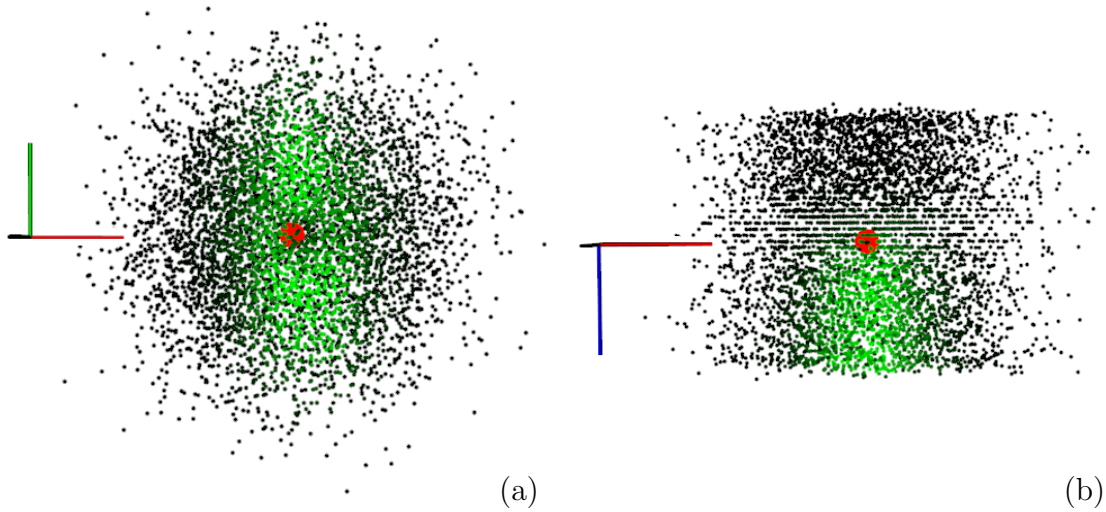
$$I(S) = -\sum_{c} p(c|S) \log p(c|S), \tag{4.9}$$

where $p(c|S)$ is the normalized distribution of the set of body part classes $c(u)$ for all $u \in S$. We normalize ig, so that larger values represent more meaningful features and lower values less meaningful features.

4.3.4. Kinematic Chain Constraints

In certain scenarios we find misclassified regions, e.g., estimated hand regions on the knees, which are easily identifiable for humans using prior knowledge about the human body.

However, random ferns independently classify each pixel w.r.t. the classified outcome of neighboring pixel classes. For this reason, we incorporate prior information about the connections of human body parts in a post-processing step. On the depth image labeled with body part estimates, we build a kinematic tree, starting from the root node at the body center. Connected pixels belonging to the same body part form a cluster. We search the shortest path to the root node for each cluster. The path length is determined by how many cluster borders are passed. Connections to clusters that do not obey the kinematic chain constraints are penalized by additional path cost. If the path is not equal to the expected path according to the kinematic chain of the skeleton, that cluster is considered a misclassification. Pixel probabilities



(a) (b)

Figure 4.11.: Randomly generated feature set, evaluated on infant training data. Gaussian $\sigma = 20$ cm for pixel offsets ϕ. Depth thresholds τ range from -20 cm to +20 cm. Pixel offsets are plotted on x (red) and y (green) axis, depth threshold on z (blue) axis. Color represents information gain, green depicts large gain (more meaningful), black small gain (less meaningful). The red dot denotes the input pixel to which the offsets are related. (a) 5000 randomly sampled features. View on XY-axes. (b) Same as (a), view on XZ-axes.

are multiplied by a reweighting factor based on the last correct neighbor in direction of the root node. The pixel labels are then updated with the most probable body part class from the reweighted probabilities. This way we get an estimate that conforms to the kinematic chain.

4.3.5. Rotation Invariance

The binary depth features used in the ferns are not invariant to rotations. Instead of trying to capture all possible variations of rotation in the training data to make the ferns rotation invariant, we train on upright positions with limited range of rotation.

In our settings, the infants are always filmed from above, so that the main body axis is displayed vertically in the camera image. If the main axis strongly diverges from being vertical, the prediction will be distorted. To solve this issue, we use the pixels of estimated body parts that belong to the trunk from the previous frame as input for principal component analysis. We rotate the feature offsets to conform to the axes we get from the first two eigenvectors of PCA. This way the classification is performed as if an upright body was given as input. An example is displayed in Fig. 4.12.

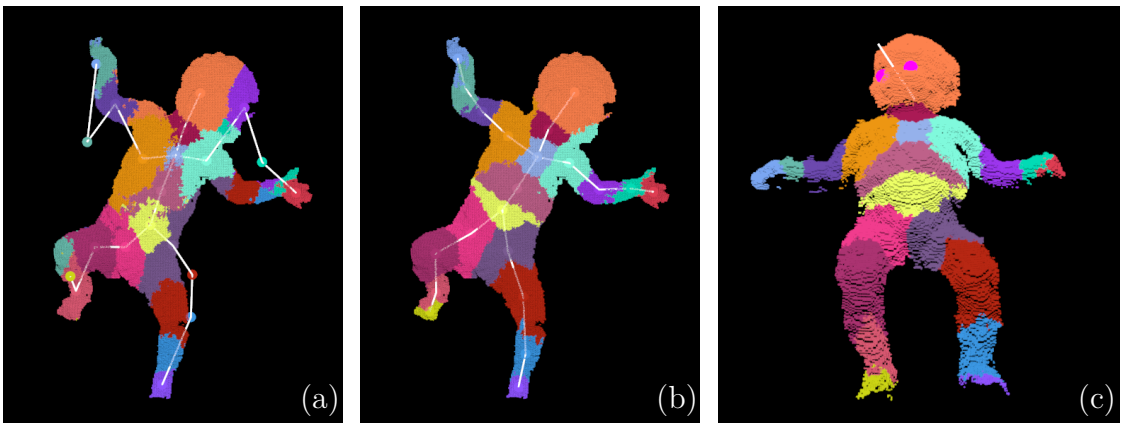


Figure 4.12.: (a) Prediction without PCA rotation correction. (b) Prediction with PCA rotation correction. (c) Estimated head rotation and pixel-wise body part labels. Pink dots depict eye positions, white line the viewing direction.

4.3.6. Head Rotation

Body joint positions do not include information about rotations around longitudinal axes, e.g., head rotation. Since the head rotation seems to be of interest for medical assessment, we calculate it based on the eye positions, which we detect in the RGB image and project to the depth image.

We use a convolutional neural network from C++ library dlib [Kin09; Kin15] to detect the infants' face in each 2D RGB image. We use a shape predictor [KS14] from the same library to find 68 facial landmarks in the detected face region, from which we extract the centers of the eyes. We project the 2D eye coordinates from the registered RGB images to the 3D point cloud constructed from the depth image. For calculating the head rotation angle, we first create a plane that is defined by the head joint and the main body axis, and is perpendicular to the table plane. The head rotation is then given as the angle between this plane and the line through the center between the eyes and the detected head joint position. And example showing the eye positions and the "viewing direction" is given in Fig. 4.12 (c).

4.3.7. Evaluation

Manual annotation of ground truth 3D joint positions is a cumbersome, yet inaccurate process. For this reason, we fit a 3D body model, *SMPL*$_{\mathrm{B}}$ *(prel.)*[1] to the recorded sequences and visually verify the plausibility and accuracy of results and consider them ground truth for our evaluation. A fixed offset is added to all predicted joint positions to account for differences between the model used for generating the

[1]This was a preliminary version of the initial body model SMPL$_{\mathrm{B}}$ from Sec. 5.3, since at the time of writing the final SMIL model, cf. chapter 5, was still work in progress.

Table 4.4.: Average joint position error (and standard deviation) in cm over all three sequences of FernSeq2. *Random ferns* denotes the approach from Sec. 4.2, *Ferns extension* uses the new training set (Sec. 4.3.2) and applies rotation correction (Sec. 4.3.5). *RWF* denotes kinematic chain reweighting and filtering (Sec. 4.3.4), and *FS* the feature selection method (Sec. 4.3.3).

Method	Without RWF	With RWF
Random ferns	1.782 (3.103)	1.382 (1.459)
Ferns extension	1.212 (0.931)	1.224 (0.897)
Ferns extension + FS	1.236 (1.121)	1.222 (0.852)

ground truth and the model for training the ferns. We compare our methods to the baseline fern approach from Sec. 4.2.3 on three recordings, which we denote as data set *FernSeq2*. Two of them are taken from longer sequences and contain 500 frames where the infants are most active. The third sequence consists of 4500 frames. One of the short sequences was recorded with a Kinect V2, the others with Kinect V1. The background is removed prior to evaluation so that only pixels are contained that are part of the infant.

Error metrics. Pose estimation approaches are mostly evaluated by indicating the *average joint position error* (AJPE). For the task of motion analysis, though, more strict evaluation measures are needed. Therefore, we use the *worst-case accuracy* (WCA) as proposed in [Tay+12], which is the percentage of frames in which *all* joints lie within a certain threshold distance from the ground truth. Additionally, we introduce a measure that we call the *jitter accuracy* (JA) in order to quantify the smoothness of joint predictions. Joint estimates that are "jumping" around the ground truth at a fixed distance will have a constant joint position error, but most probably would harm the motion analysis stage. We calculate the difference of predicted joint position deviation relative to ground truth in consecutive frames (*jitter error*), which we define by

$$je_{i,j} = \|(x_{i,j} - gt_{i,j}) - (x_{i-1,j} - gt_{i-1,j})\|, \tag{4.10}$$

where $i \in \{2, \ldots, N\}$ is the frame number, j the joint index, $x_{i,j}$ the predicted position of joint j in frame i, $gt_{i,j}$ the ground truth position of joint j in frame i and $\|\cdot\|$ the Euclidean norm. The jitter accuracy is the percentage of frames in which the jitter error of *all* joints is smaller than a certain threshold.

Results. We compare the proposed methods to the baseline approach (Sec. 4.2.3). All classifiers are trained using the same number of ferns (15), fern depth (13) and pixel offset neighborhood radius (20 cm). PCA rotation correction is applied with all methods. *Ours* stands for the proposed approach, trained on the new training

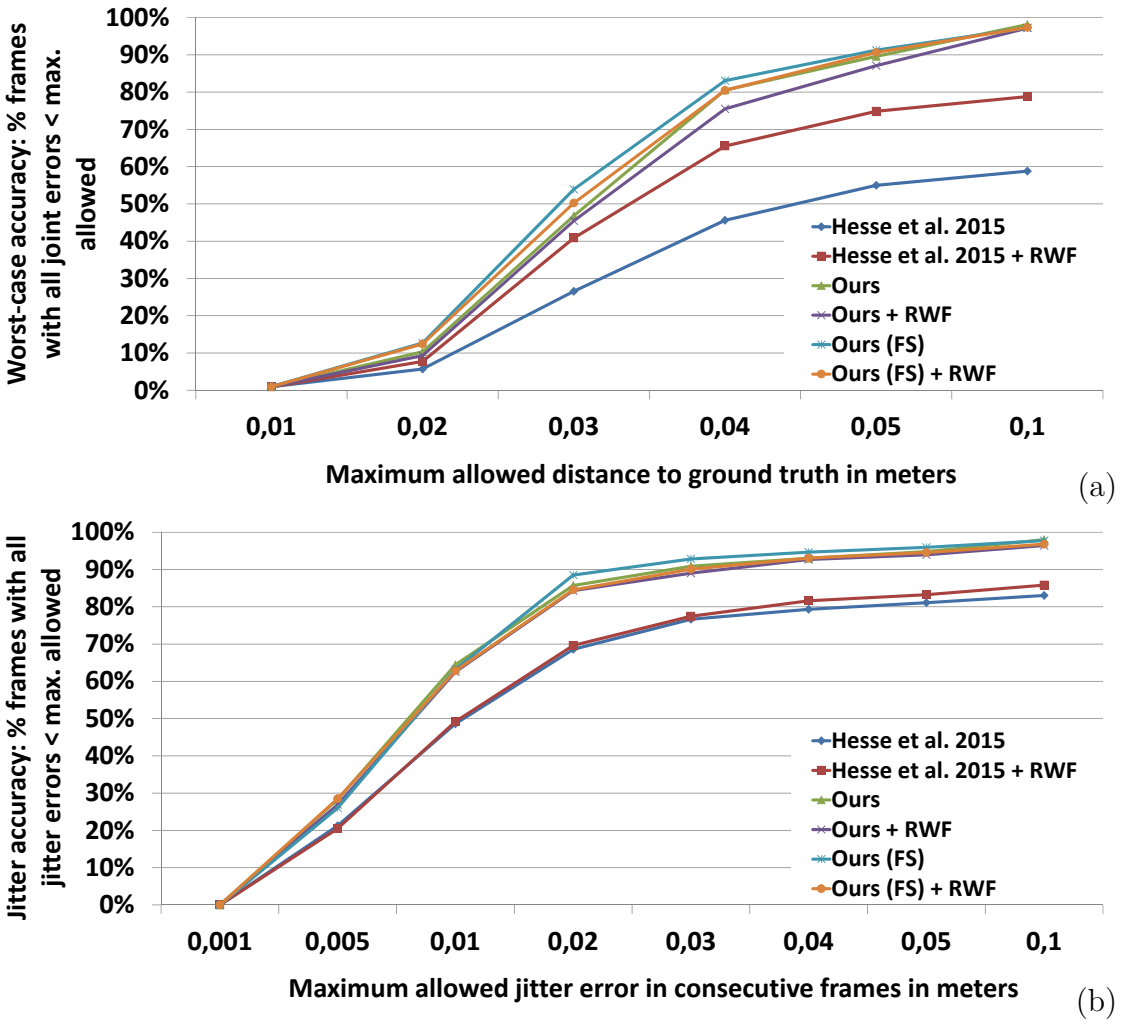


Figure 4.13.: Evaluation results for FernSeq2. (a) Worst-case accuracy. (b) Jitter accuracy. Average over all sequences.

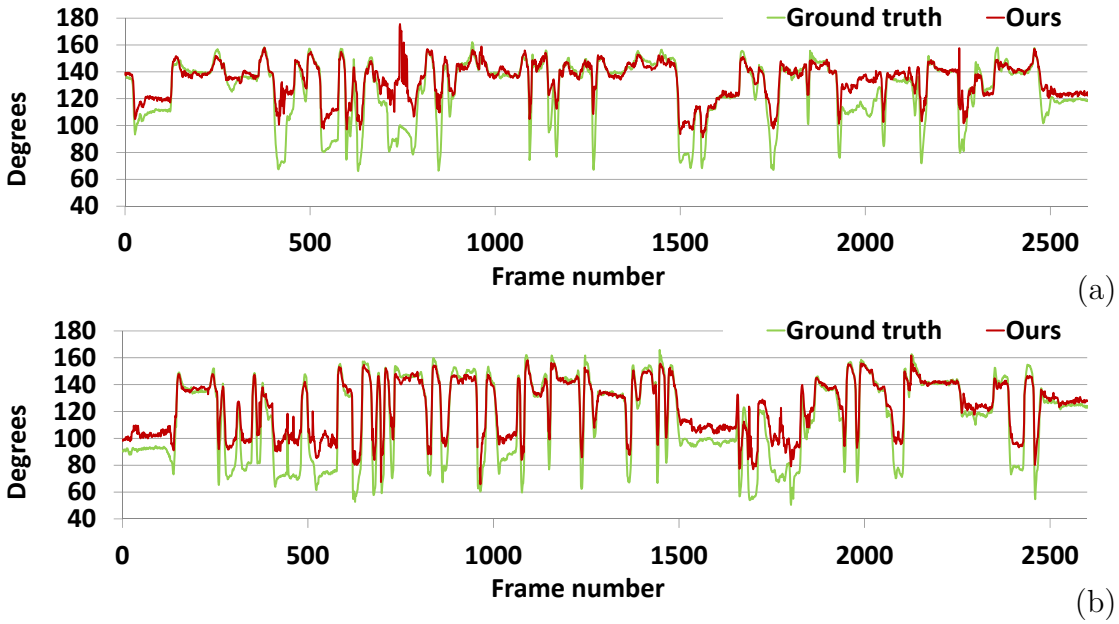


Figure 4.14.: Comparison of estimated and ground truth knee angles on 2600 frames from the longest sequence of FernSeq2. (a) Left knee angles. (b) Right knee angles.

set, *FS* indicates that the feature selection step was used prior to training, *RWF* means that reweighting and filtering is applied.

Results of the evaluation are displayed in Tab. 4.4, and Fig. 4.13. Worst-case accuracy is improved by roughly 10% ($maxdist \geq 3cm$) over the baseline if the proposed reweighting and filtering step is applied. AJPE is reduced by 0.4 cm, while the jitter accuracy remains at a similar level.

When training the approach using our new baby-like poses (*Ours*), accuracy increases between 10 and 20% ($maxdist \geq 3cm$) compared to [Hes+15], while reducing AJPE by more than 0.5 cm. Standard deviation of AJPE is reduced marginally by applying RWF to *Ours*, while the average error slightly increases. From our experiments, we find the explanation that RWF fixes large misplacements of body parts, e.g., knee classified as hand - hence the improvement over [Hes+15], but due to the fact that class labels of body part patches are changed, predicted positions seem to shift too far sometimes. The feature selection step does not lead to a big improvement in accuracy (*Ours* vs. *Ours (FS)*). If there are no redundant features included in training without feature selection, there will be no benefit from this step. We still believe feature selection to be a valid enhancement, because it will prevent the inclusion of redundant features independent of the amount of randomness used during training.

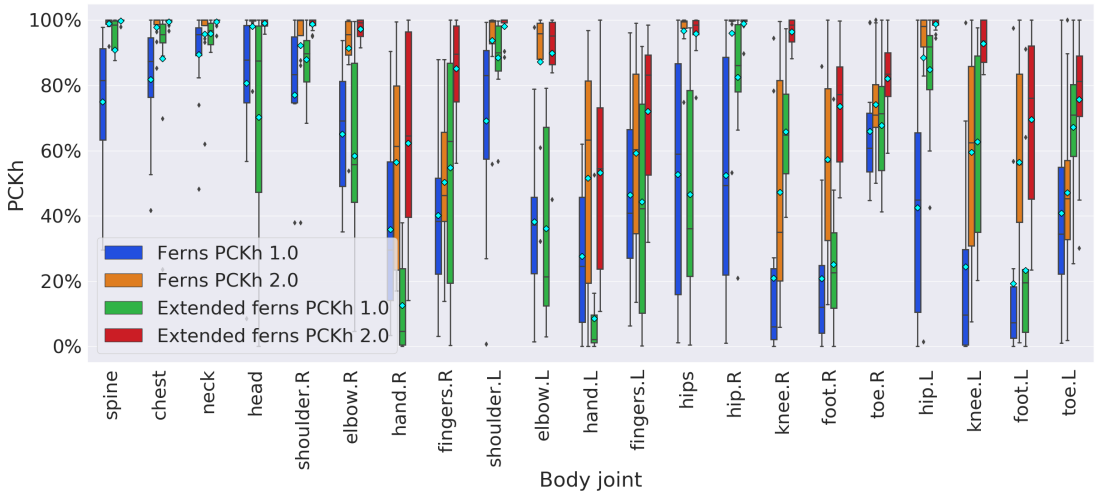


Figure 4.15.: MINI-RGBD results for 3D pose estimation based on random ferns and extensions. Percentage of correct keypoints in relation to head segment length (PCKh) per joint. Cyan diamonds depict the mean value.

Angle comparison. In Fig. 4.14, we illustrate how our system captures movement information like joint angles accurately by comparing it to angles calculated from ground truth joint positions. Results are presented for left and right knee for 2600 frames of the longest sequence of FernSeq2. Although the angle values do not match the peak levels exactly, we observe that the estimated angles reflect the ground truth very well. Doctors will be able to get a first impression of the movement quality by one glance on the plotted angles. It will, e.g., be clearly visible if there is an absence of motion on one side of the body.

Evaluation on MINI-RGBD data set. Similar to the baseline ferns approach, we evaluate the extended approach on the MINI-RGBD data set using the same metrics.

We present results in Fig. 4.15 and Fig. 4.16. Mean average precision – i.e., average PCKh over all joints and sequences – for PCKh 1.0 is 64.2%, and 91.0% for PCKh 2.0.

Very high PCKh 2.0 rates are achieved for torso and head body parts, while lowest rates are obtained for joints related to extremities (Fig. 4.15). PCKh 1.0 rates differ a lot from PCKh 2.0 for elbows, hands, and feet. We observe that the estimated hand and foot regions are often too large, leading to the hand joints lying more in the direction of the elbow, respectively the foot joints in direction of the knees. With an expansion of the threshold for correctness (PCKh 2.0) these displacements are accepted as correct, leading to large jumps from around 30% (PCKh 1.0) to 70 - 80% (PCKh 2.0).

The average joint position error (AJPE) over all sequences and joints is 2.86 cm,

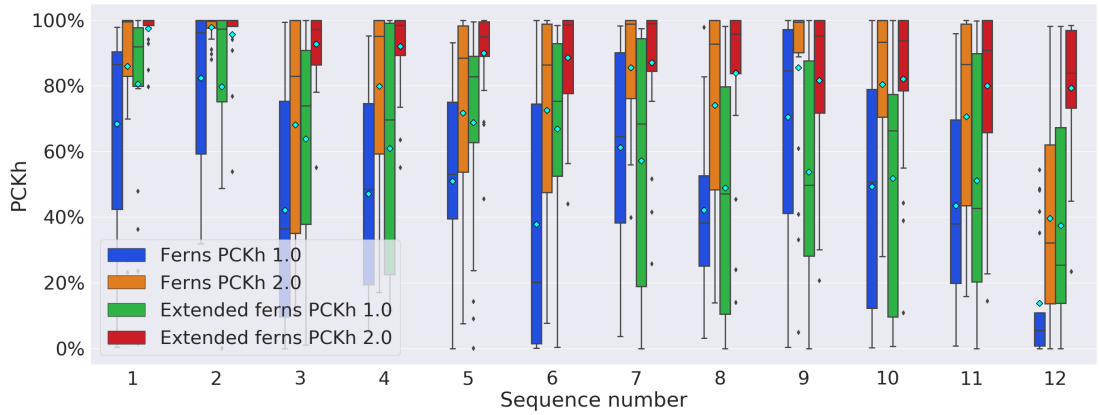


Figure 4.16.: MINI-RGBD results for 3D pose estimation based on random ferns and extensions. PCKh per sequence. Cyan diamonds depict the mean value.

compared to 4.66 cm of the baseline. Joint position errors are largest for the extremities, at an average distance to ground truth of up to 5 cm (Fig. 4.17 (a)).

If the estimate for a joint was missing in a frame – no pixel was labeled with that joint's label – we ignored this joint for the calculation of AJPE, i.e., we only divided the sum of joint errors by the number of actually estimated joints. The number of frames with missing estimates, denoted by joint (in 12K frames, average for left and right sides): neck 37, elbows 12, hands 62, fingers 156, feet 79, toes 607, all others 0. For the calculation of PCKh metric, missing joints were considered as lying outside the correctness threshold.

The evaluated approach shows high accuracy when arms and legs are moving beside the body, but the accuracy decreases, especially for hands and feet, when limbs move close to or in front of the body. This becomes extremely visible in sequence 9, where the infant moves the left arm to the right side of the body multiple times, leading to the highest overall AJPE of 4.7 cm (Fig. 4.18). Best AJPE results are achieved for sequence 2, at 1.46 cm, which is close to results of the evaluation of three recordings above. The varying accuracy for different sequences shows the levels of difficulty and the variance of motion patterns included in the data set.

Altogether, we observe that the proposed extensions outperform the baseline by a large margin on multiple error metrics.

4.4. Discussion and Limitations

We proposed a method for estimating 3D body joint positions in depth images. The training time is reduced by nearly two orders of magnitude compared to a similar approach used in commercial applications.

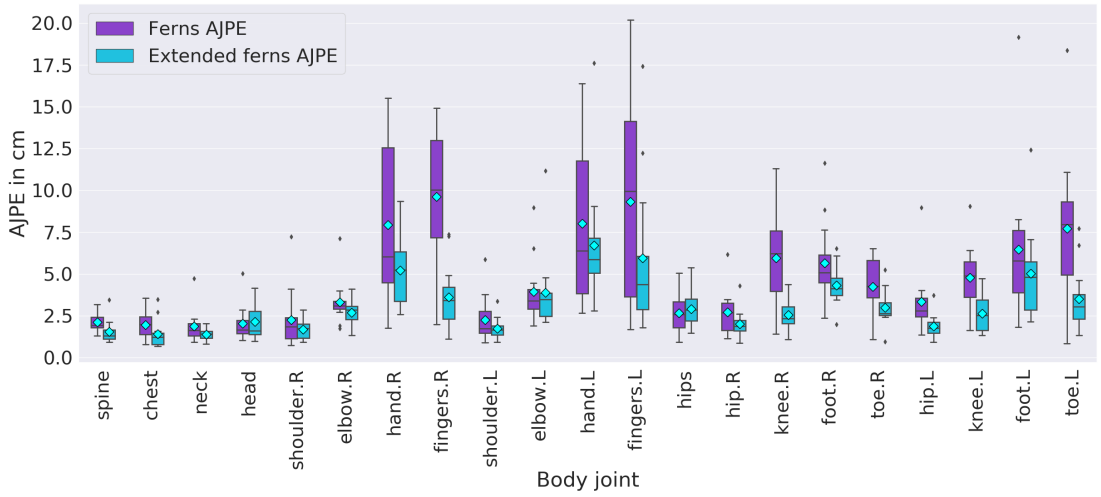


Figure 4.17.: MINI-RGBD results for 3D pose estimation based on random ferns and extensions. Average joint position error (AJPE) per joint. Cyan diamonds depict the mean value.

However, several limitations exist. We observe a degradation of accuracy in the presence of occlusions, or if hands are moved close to the body. Our approach, similar to [Sho+11], processes one image at a time and does not take into account temporal information (except in our rotation invariance extension where the estimate of the previous frame is considered). On the upside, working on single images prevents the method from getting stuck in local minima. Yet, it does not necessarily provide smooth and consistent estimates for consecutive frames, and by incorporating a tracking procedure, the handling of failure cases and overall accuracy could be improved.

The system was trained either on adult poses that differ from infant poses, or on solely synthetic poses by manually defining angle ranges and sampling inside these ranges. These most certainly do not include all possible valid body configurations an infant can produce. With the model-based method we will present in the next chapter, it would be possible to generate larger amounts of training data, as described in chapter 6. At the time of writing, we have not yet generated a training set large enough to re-train the random ferns approach.

We only use RGB information when detecting the eyes for capturing head rotation. Instead, the RGB information, which is complementary to depth, could be used to further improve accuracy, e.g., by learning appearance models of different body parts as in [Buy+14].

A general drawback of the proposed approach is that the estimated joint positions do not capture body part rotations. It is unclear how important these rotations are for medical assessment, but what *is* clear is that information is lost.

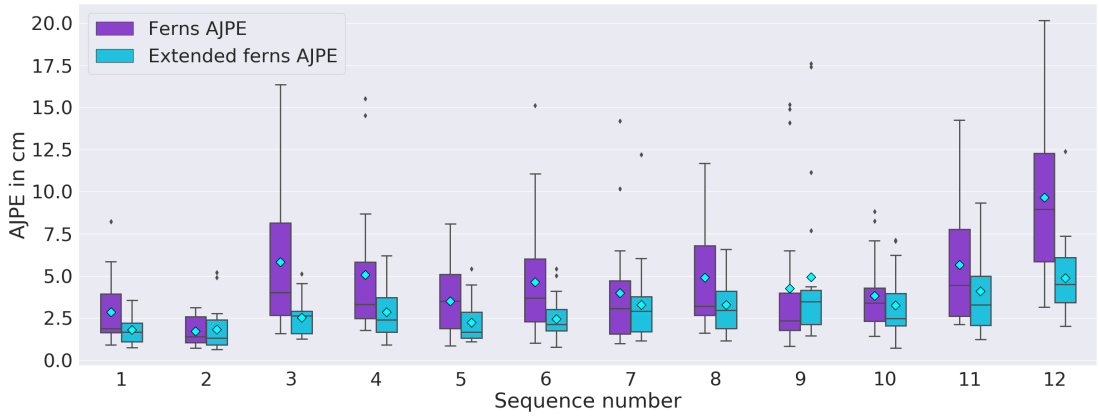


Figure 4.18.: MINI-RGBD results for 3D pose estimation based on random ferns and extensions. AJPE per sequence. Cyan diamonds depict the mean value.

5. Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D sequences

In the previous chapter, we discussed a fast and easy to train approach for pose estimation from depth images. We identified several limitations, which motivates the development of a more advanced method in this chapter.

The general movement assessment judges the movement quality from a holistic view of movement complexity and variation of all parts of the body. Following this concept, a 3D body model seems like a good choice for representing motion, as it allows reasoning about the complete body surface as well as the underlying skeleton.

Researchers have shown that it is possible to accurately register body models to point cloud data (e.g., generated from depth images). Parameters related to motion, like joint angles and positions, can be easily extracted from such registrations.

In this chapter, we will detail our approach to model-based estimation of infant shape and pose from RGB-D data. First, we establish our notion of the term *body model*, before we give an overview of methods for data-driven body model creation and how they have been used for capturing motion. Because of the fact that no realistic infant body model exists, we propose a method to learn an infant-specific model. Opposed to learned adult models – which generally leverage large data sets of high-quality 3D scans – no such repository exists for infants, which is why we learn our model from low-quality, incomplete RGB-D data. Finally, we show in our experiments that the learned model captures sufficient motion detail for infant motion analysis (GMA).

5.1. Statistical Human Body Models

When using the term "body model", we refer to a surface representation of the body. Such model is parameterized with the shape, giving information about the joint locations, and the pose, defining the angles between the limbs. Statistical body models are learned from data, opposed to artist-created models, which generally are not able to achieve the same level of realism.

Human body models are commonly represented as a set of body parts, connected by a kinematic chain. A change in position/rotation of a body part results in a change

of position of "children" body parts. Usually, pose is described by a vector of joint angles. To summarize, the model is a mapping from shape and pose parameters to vertices.

5.1.1. Basics of Parametric Models

In this section, we give an overview of the creation of a posed and shaped body mesh from model parameters.

We follow the formulation of Pons-Moll and Rosenhahn [PMR11], and Murray [Mur94].

[PMR11] discuss possible choices of parameters and define desirable properties of a model parameterization for human motion as follows:

- Pose configurations are represented with the minimum number of parameters.
- Human motion constraints, such as articulated motion, are naturally described.
- Singularities can be avoided during optimization.
- Easy computation of derivatives of segment positions and orientations w.r.t. the parameters.
- Simple rules for concatenating motions.

We examine the *axis-angle* representation for mapping a vector of *pose* parameters (joint angles) to an output mesh, given a body model with underlying kinematic chain. Different representations like Euler angles and quaternions are discussed in [PMR11], but violate one or more of the desirable properties, which is why we focus on the axis-angle representation.

The human body has different kinds of rotational joints. For some of them it is necessary to define a fixed axis around which rotations are possible, e.g., the knee. For notational simplicity, we assume rigid body segments, i.e., the distance between any two points of a body segment remains constant. A compact way to represent this axis-angle representation with three parameters is to define the rotation around a unit length axis vector ω by an angle θ as $\theta\omega$.

In order to calculate the movements of connected body segments with the axis-angle representation we need to apply some transformations.

We can create a rotation matrix R that defines the rotation of θ around ω by using the exponential form

$$R = \exp(\theta\hat{\omega}), \tag{5.1}$$

with $\hat{\omega} \in \mathrm{so}(3)$ being a skew symmetric matrix constructed from ω with so(3) denoting the set of skew symmetric matrices that hold for $\{A \in \mathbb{R}^{3\times3} | A = -A^T\}$ [Mur94].

The skew symmetric matrix $\theta\hat{\omega}$ for $\omega = (\omega_1, \omega_2, \omega_3)$ is calculated as

$$\theta\hat{\omega} = \theta \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \tag{5.2}$$

The multiplication of the matrix $\hat{\omega}$ with a point p is equivalent to the cross product of the vector ω with p.

The exponential of $\theta\hat{\omega}$ can be calculated using the Rodriguez formula:

$$\exp(\theta\hat{\omega}) = I + \hat{\omega}\sin(\theta) + \hat{\omega}^2(1 - \cos(\theta)). \tag{5.3}$$

The advantage is that only the square of $\hat{\omega}$, as well as sine and cosine of real numbers have to be computed. The construction of a rotation matrix from angle and axis by using simple operations is a favorable property and promotes the use of the axis-angle representation.

In the next step, the formulation will be extended to represent rotation and translation, i.e., the exponential map for rigid body motions. A twist $\xi = (\upsilon, \omega)$ contains translation parameters $\upsilon = (\upsilon_1, \upsilon_2, \upsilon_3)$ and rotation parameters ω as above. The translation is along the axis of rotation, starting at the location of the axis.

The exponential of any rigid motion $G \in \mathbb{R}^{4\times4}$ is given by

$$G(\theta, \xi) = \exp(\theta\hat{\xi}). \tag{5.4}$$

$\theta\hat{\xi}$, a 4×4 matrix, is called the *twist action* [PMR11] and is a generalization of 5.2. The twist action can be constructed from the twist coordinates $\theta\xi \in \mathbb{R}^6$ as

$$\theta\hat{\xi} = \theta \begin{pmatrix} 0 & -\omega_3 & \omega_2 & \upsilon_1 \\ \omega_3 & 0 & -\omega_1 & \upsilon_2 \\ -\omega_2 & \omega_1 & 0 & \upsilon_3 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.5}$$

and its exponential is given as

$$\exp(\theta\hat{\xi}) = \begin{pmatrix} \exp(\theta\hat{\omega}) & (I - (\exp(\theta\hat{\omega})(\omega \times \upsilon + \omega\omega^T\upsilon\theta) \\ 0_{1\times3} & 1 \end{pmatrix}, \tag{5.6}$$

with $\exp(\theta\hat{\omega})$ as before.

Forward kinematics. Finally, we can transform a point on a body segment from body coordinates p_b to world (spatial) coordinates p_s by applying joint rotations along the kinematic chain towards the body segment. This is achieved by concatenating the transformation matrices of joints in the kinematic chain.

The point in spatial coordinates can then be calculated as

$$\bar{p}_s = G_{sb}\bar{p}_b, \tag{5.7}$$

with $\bar{p}$ denoting point p in homogeneous coordinates, i.e., $\bar{p} = [p, 0]$.

The transformation between spatial and body frames is given by

$$G_{sb}(\Theta) = e^{\hat{\xi_1}\theta_1} e^{\hat{\xi_2}\theta_2} \cdots e^{\hat{\xi_n}\theta_n} G_{sb}(0), \tag{5.8}$$

where $\Theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ is the joint angle vector, and ξ are constant twists in the reference configuration, which is the zero pose. $G_{sb}(0)$ is usually chosen to be the identity.

The formulation considering purely rigid body parts leads to discontinuities and artifacts close to joints when the joint angles deviate from zero. The standard solution for resolving these discontinuities and to smooth shape at body joints is called *linear blend skinning* (LBS) . LBS assigns vertices to multiple body parts and using a weighted linear combination of the influence of these parts.

Eq. 5.7 is then transformed to

$$\bar{p}_s = \sum_{k=1}^{K} w_k G_{sb}\bar{p}_b, \tag{5.9}$$

with K denoting the number of joints, and w_k the weight of joint k for point $\bar{p}_b$. LBS solves the problem of discontinuity, but often does not lead to visually realistic results. Multiple methods have been proposed to improve realism, e.g., the SMPL model – which we review in the next section – applies corrective blend shapes.

After laying the mathematical foundations of how joint angle transformations are applied to model vertices, we describe the specific body model on which our approach in Sec. 5.3 is based.

5.1.2. Properties of the Skinned Multi-Person Linear Body Model

The *Skinned Multi-Person Linear model* (SMPL) is a statistical body model that was learned from thousands of high quality 3D scans [Lop+15]. It consists of 6890 vertices and produces a watertight body mesh from pose and shape parameters. It has a mean template shape $\bar{T}$, and the zero pose (all pose parameters equal to zero) is the so called T-pose, with fully extended arms and legs. The body shape is decomposed into *person-specific* shape and *pose-dependent* shape, i.e., non-rigid body deformations caused by pose.

The *blend shapes* describe these shape deformations, and are modeled as a vector of offsets from the model vertices in rest pose. Person-specific shape is represented by

a linear function B_s in which each orthonormal principal component of 3D shape displacements S_n is multiplied by the corresponding shape parameter β_n.

$$B_s(\beta;\mathcal{S}) = \sum_{n=1}^{|\beta|} \beta_n S_n \tag{5.10}$$

with $\mathcal{S} = [S_1, \ldots, S_{|\beta|}]$, and β denoting the full set of shape parameters.

The pose parameters θ are expressed in axis-angle representation and contain one value per degree of freedom of the K body joints, plus three parameters for the orientation of the root joint. With $K = 23$, this results in $23 * 3 + 3 = 72$ pose parameters. The pose-dependent blend shapes B_P are defined as

$$B_P(\theta;\mathcal{P}) = \sum_{n=1}^{9K} (R_n(\theta) - R_n(\theta^*))P_n, \tag{5.11}$$

where $R(\theta) : \mathbb{R}^{|\theta|} \mapsto \mathbb{R}^{9K}$, is a function that maps pose parameters to a vector of concatenated rotation matrices, and R_n denotes the nth element of R. θ^* denotes the rest pose, and $\mathcal{P} = [P_1, \cdots, P_{9K}]$ vertex offsets similar to $\mathcal{S}$.

The blend shapes B_S and B_P are added to the mean template in zero pose $\bar{T}$:

$$T_P(\beta, \theta) = \bar{T} + B_S(\beta) + B_P(\theta) \tag{5.12}$$

This results in a mesh, still in rest pose, containing the pose-dependent shape deformations as well as the identity-dependent shape information. This mesh is then posed with a standard blend skinning function W (cf. Sec. 5.1.1). The final model is then

$$M(\beta, \theta) = W(T_P(\beta, \theta), J(\beta), \theta, \mathcal{W}). \tag{5.13}$$

with $J(\beta)$ denoting a function to predict joint positions, and the blend weights $\mathcal{W}$ which encode the effect of different joints on each vertex in calculation of the skinning function.

By changing shape parameters, a weighted combination of shape blend shapes (vertex offsets) is added to the mean template vertices to produce a new shape. By changing pose parameters, body parts are rotated according to angle values (from pose parameters) using a standard blend skinning function (e.g., linear blend skinning), and corrective pose blend shapes (again, vertex offsets) are added to the model vertices to account for the pose-dependent soft-tissue deformations and errors caused by the skinning function.

The *training* of SMPL is conducted separately for pose and shape parameters. The data sets for training contain meshes that have been aligned to high-quality 3D scans. The shape data set consists of 3800 subjects in roughly the same pose from the CAESAR data set [Rob+02], and the pose data set contains nearly 1800 registrations

of 40 subjects in different poses. The model parameters are trained to minimize the distance between model and registrations.

The *shape space*, spanned by the shape parameters, consists of the mean shape and the principal shape directions. To perform principal component analysis (PCA) for extracting on a set of body meshes, these need to be in the same pose to ensure a separate modeling of shape and pose. This is achieved by first estimating the pose of a given subject mesh, and then to normalize the pose, i.e., transforming it to the zero pose. In section Sec. 5.3.6, we follow the same approach to learn our infant-specific shape space.

We further learn a prior on the plausible poses for our infant model. Such *pose prior* aims to describe the valid range of joint angles. Angle limits depend on pose, which is why learning them from a large set of different poses should be preferred over the calculation of joint limits in a fixed pose [AB15]. A pose prior is often used to constrain the estimation of human pose from images. Our prior P_{pose} consists of mean μ and covariance matrix Σ, and determines the validity of given pose parameters θ by calculating the Mahalanobis distance

$$d(P_{\text{pose}}, \theta) = \sqrt{(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)}. \tag{5.14}$$

This section has described the mathematical foundations of how a body model is transformed with respect to pose and shape parameters. In the next section, we review the literature on the creation of body models, as well as their application to pose and shape estimation from (depth) images.

5.2. Related Work – Shape and Pose Estimation using Body Models

First, we give an overview of methods for the data-driven creation of human body models. Then, we show how body models are used in literature for capturing human pose and shape.

Model creation. Statistical parametric models of the human body surface are usually based on the *Morphable Model* idea [BV99], stating that a single surface representation can morph and explain the different samples in a population. These models can be intuitively viewed as a mathematical function taking shape and pose parameters as input and returning a surface mesh as output.

One option to create such model is to have an artist sculpt a body mesh together with a manual definition of possible deformations. This approach often lacks realness, as it is impossible to include all details and dependencies of real human shape and pose. Therefore, the preferred solution has been to learn the statistics of human bodies from data.

The *shape space* and the *pose prior*, i.e., the range of most plausible poses, are learned by registering the single surface representation, i.e., a template surface, to real-world data. The shape space and the pose prior allow a compact representation of the human body surface describing the geometry of the human body surface of an observed population in a low-dimensional space.

Existing models have been learned from different real-world data. For example, to model the variation of faces, De-Carlo et al. [DeC+98] learn a model from a cohort of anthropometric measurements. Blanz and Vetter, use dense geometry and color data to learn their face model [BV99]. Allen et al. use the CAESAR dataset to create *the space of human body shapes* by using the geometry information as well as sparse landmarks identified on the bodies [All+03]. Similarly, Seo and Magnenat-Thalmann learn a shape space from high quality range data [SMT03]. Angelov et al. propose SCAPE, a statistical body model learned from high quality scan data, which does not only contain the shape space, but also accounts for the pose dependent deformations, i.e., the surface deformations that a body undergoes when different poses are taken [Ang+05]. SCAPE and all successive statistical models of the human body surface [Has+09; Hir+12; FB12; Che+13; ZB15; Lop+15; Pis+17] or their soft tissue dynamics [PM+15a; Lop+15; Kim+17] have been learned from a relatively large number of high quality range scans of adult subjects. Adults, in contrast to infants, are typically cooperative and can be instructed to strike specific poses during scanning.

Animal shape modeling methods face a similar difficulty as ours: live animals are generally difficult to instruct and their motions make them difficult to scan. Thus, they provide a source of inspiration to create models without a large cohort of high quality scans. Cashman and Fitzgibbon learn a deformable model of dolphins by using manually annotated 2D images [CF13]. Kanazawa et al. learn the deformations and the stiffness of the parts of a 3D mesh template of an animal from manually annotated 2D images [Kan+16]. In [Kan+18b], they learn category-specific textured 3D meshes from collections of 2D images. To create the animal model SMAL, Zuffi et al. circumvent the difficulty to instruct and scan real animals by using a small dataset of high quality scans of toy figurines [Zuf+17]. The SMAL model can be fit to new animals using a set of multi-view images with landmarks and silhouette annotations.

Capturing motion using a body model. The existing body models have proven to be successful in capturing the pose and shape of a subject from different types of input data. Registration aims at explaining the observation (input data) with the model by minimizing an objective function that measures the difference between the two. Often, this includes the distance between model and data points, but also additional terms like prior probabilities on pose and shape to avoid the model to deform unnaturally to explain the data. Once a model is registered to the input data, one can obtain the desired motion information from the registration. As the

joint locations depend on the shape, the closer the estimated body shape is to the actual subject's shape, the better, i.e., more accurate, the tracking of motions will be.

Shape and pose estimation from RGB-D. RGB-D data provides information about the scene in 3D, thus allowing to infer 3D shape information and to resolve ambiguities that occur in RGB images, i.e., different poses producing the same image.

Researchers have used different models for different use cases. Rather simplistic models are used for a maximum of processing speed, personalized avatars are used for realistic appearance, and general generative models for less constrained scenarios.

In models that are composed of **simplistic shapes**, shape estimation refers to changes of size or diameter of the shapes instead of estimating realistic human shapes. Such adaptations contribute to a better description of input data by the model surface.

Knoop et al. apply an articulated cylinder model with elastic band joints to perform close to real-time motion capture from depth data [Kno+06]. They include information about 2D points detected on the human body to constrain the 3D fitting process, similar to our method for initialization (Sec. 5.3.4).

Ganapathi et al. assemble a model from 3D capsules, and adapt their overall scale to input data. Their iterative closest point method incorporates free-space and self-intersection constraints and runs in real time [Gan+12].

One such simplistic model has been created for infants, consisting of basic shapes like cylinders [Ols+15]. This model is aligned to RGB-D data by minimizing the distance between model and data without further constraints like a pose prior. Opposed to ours, this method only capture joint positions and not rotations of body parts due to the uniform shapes of which the model is assembled.

The generative model fitting approach has been **combined with discriminative methods** to leverage the advantages of either one.

Instead of applying global optimization with a generative model, a body part detector creates a good initial guess, so that the optimization only needs to be applied locally to achieve accurate results.

[Baa+13] have created a pose data base using a commercial marker-based motion capture system. Their discriminative tracker extract sparse depth features, i.e., geodesic extrema, which are used to find the best match in the pose data base. Both, this candidate match, and a candidate generated from the previous frame, are locally optimized to fit the data. The resulting pose hypotheses are fused in a voting scheme to give the final result. This real-time approach uses a fixed size model and scales input data to match the model size.

Mode detailed parametric body models generally model the space of human bodies without clothing or hair. Some approaches propose to overcome the discrepancy between such a shape space and the real world by creating personalized avatars from the input data and then registering them to the dynamic sequences by keeping the personalized shape fixed. In this context, we can distinguish between i) creating a personalized *mesh*, possibly created from one or more scans, and attaching it to a generic template that can be reposed, and ii) fitting a parametric model to one or multiple scans to describe the subject's shape, which will then lie in the model's shape space. The former may give more detailed and more realistic looking results by including a texture map, as well as hair and clothing shape. The dynamics of these parts, however, are not included in the model. The latter approach is applicable in more general scenarios, e.g., where subjects can not take predefined poses and no prior information about the subject is given.

The limitation of a single monocular camera producing incomplete data is overcome by fusing multiple RGB-D scans of a person into a single reference frame to produce something more similar to a full 3D scan. This creates an additional step and usually requires cooperative subjects to take predefined poses, which makes it unfeasible for infants.

Creating personalized models from RGB-D. The creation of personalized models is performed by capturing the detailed shape and appearance and then binding it to a generic template. This personalized template can then be posed according to the kinematic chain of the underlying template. Tong et al. register multiple partial Kinect scans to form a complete body mesh that can be animated and posed [Ton+12]. Cui et al. create textured 3D meshes by registering Kinect scans of subjects rotating 360 degrees in a static T-pose [Cui+13]. Zhang et al. merge multiple *Kinect fusion* scans in different poses from different viewpoints and create personalized dynamic models based on a generic template model [Zha+14]. Shapiro et al. create a personalized avatar from 4 static scans at 90 degree intervals. The surface is bound to a skeleton, which can be animated with new poses [Sha+14]. Perbet et al. estimate body shapes under clothing from depth images by constraining the model to lie inside the 3D point cloud [Per+14].

Fitting personalized models to RGB-D data. Instead of a fixed model, Helten et al. create a personalized template by fitting a (simplified) parametric body model (SCAPE) to two Kinect scans, front and back [Hel+13a]. They adapt the pose data base of [Baa+13] to the shape of their template, and perform the same tracking procedure with the personalized template. The objective function minimizes the distances between 3D point cloud and model surface. Subsequent work shows that an increase in accuracy can be achieved by integrating IMU measurements [Hel+13b]. Qammaz et al. rely on the method by [Sha+14] to create personalized models, which they fit to new RGB-D sequences [Qam+18]. Similar to our initial-

ization stage (Sec. 5.3.4), they leverage estimated 2D joint positions to constrain the 3D model fitting. Since they obtain a personalized shape, they keep the shape fixed and optimize their objective function w.r.t. pose parameters. Ye et al. [Ye+12] use a laser scanner to create high quality personalized models of subjects in the scene. However, they lift constraints that are applied in other approaches, by allowing multiple subjects to interact with each other, and tolerating multiple moving (handheld) cameras. They leverage information from three Kinects for optimizing an energy function taking into account geometric correspondences, scene segmentation and image correspondences. The detailed geometry and appearance in combination with multiple camera views allows them to capture challenging interactions of multiple people. The high-quality laser scanning, however, limits the general applicability. Zheng et al. [Zhe+14] create personalized avatars based on RGB-D data from 4 cameras. They re-train a SCAPE model with an expanded human pose database to enlarge the tracking space of SCAPE. They develop a correspondence estimation method based on articulated ICP for tracking motion sequences.

Fitting generic parametric body models to RGB-D data. To avoid the preliminary step of personalized shape creation, other methods use generic body models to capture pose. Weiss et al. describe a method for fitting a body model, SCAPE, to four static RGB-D scans from different viewpoints [Wei+11]. They fit a ground plane to the scene and use the plane information to refine the model position for intialization. They overcome the problem of incomplete data by using multiple views of the subject in roughly the same pose to extract full body shape information. In their objective function, they minimize the silhouette distance and the depth distance. They fit their model to arbitrary new poses and assume a coarse initial pose estimate. They report a processing time of 65 mins per body. Ye and Yang introduce a method for real-time shape and pose tracking from depth [YY14]. They register an articulated deformation model to point clouds within a probabilistic framework. Yu et al. introduce an approach for real-time reconstruction of non-rigid surface motion from RGB-D data using skeleton information to regularize the shape deformations [Yu+17]. They extend this approach by combining a parametric body model to represent the inner body shape with a freely deformable outer surface layer capturing surface details [Yu+18].

Bogo et al. [Bog+15] fit a multi-resolution body model to Kinect sequences. This work is the closest to ours, which is why we give a brief summary and identify similarities and differences. Bogo et al. aim at creating highly realistic textured avatars from RGB-D sequences. Shape, pose and appearance of freely moving humans are captured using a parametric body model, which is learned from 1800 high-quality 3D scans of 60 adults. Shape information is accumulated over each sequence in a "fusion cloud" which allows to create detailed personalized shapes. The captured subjects wear tight clothing and take a predefined pose at the beginning of each sequence, as is common in scanning scenarios. The body model is used at different resolutions in a coarse-to-fine manner to increasingly capture more details. A dis-

placement map represents fine details that lie beyond the resolution of the model mesh. The model contains a head-specific shape space for retrieving high-resolution face appearance.

In our work, we also capture shape and pose of sequences containing unconstrained movements. To create a personalized shape, we also merge all temporal information into fusion clouds. In the gradient-based optimization, some of our energy terms are similar to the ones from [Bog+15]. In contrast to their work, an initial infant body model is not available and we must create it. We adapt an existing adult body model to a different domain, namely infants, to be used as an initial model for registering our infant sequences. The fact that infants lie in supine position in our scenario presents two different constraints. First, it means that very few backs are visible and we have to deal with large areas of missing data in our fusion clouds. Second, as infants are in contact with the background, i.e., the examination table, we can not rely on a background shot to segment the relevant point cloud. When the infants move, they wrinkle the towel they are lying on with their hands and feet. However, we can take advantage of the planar geometry to fit a plane to the table data to segment it. Moreover, we can (and do) use the fitted plane as a geometric constraint, as we know the back of the infants can not be inside the examination table. Also, in contrast to Bogo et al. [Bog+15], we can not rely on predefined poses for initialization since the infants are too young to strike poses on demand. We contribute a new automatic method for choosing the best poses for initialization. Moreover, the clothing in our setting is not constrained: we have to deal with diapers, onesies and tights. In particular, diapers pose a challenge since their shape largely deviates from the human body. We handle the unconstrained cloth condition by segmenting the points corresponding to clothes, and by introducing different energy terms for clothing and skin parts. Finally, in our work we do not use the appearance of the surface, but rather use the RGB information to extract 2D landmark estimates to have individual constraints on the face and hand rotations.

Many more model-based methods have been proposed for estimating 3D body shape and pose from different data types, namely **RGB images** [Gua+09; Has+10; Bog+16; Las+17; Zuf+18; All+17; All+18; Kan+18a; Var+18; Tun+17], **multi-view RGB images** [Sto+11; Str+12b; Str+12a; Liu+13; Hua+13; Hua+17; Elh+15; Elh+17; Rho+18], **sparse marker positions** [Lop+14], **3D scans** [All+03; Hir+11; Hir+12; Bog+14; Rom+17; Hua+18; PM+17; Zha+17; Wuh+14], and **4D sequences** (3D over time) [Bog+17].

Adults and infant bodies differ in scale and proportions, which is why simply scaling an adult body model to infant size will not yield satisfying results, as can be viewed in Fig. 5.1.

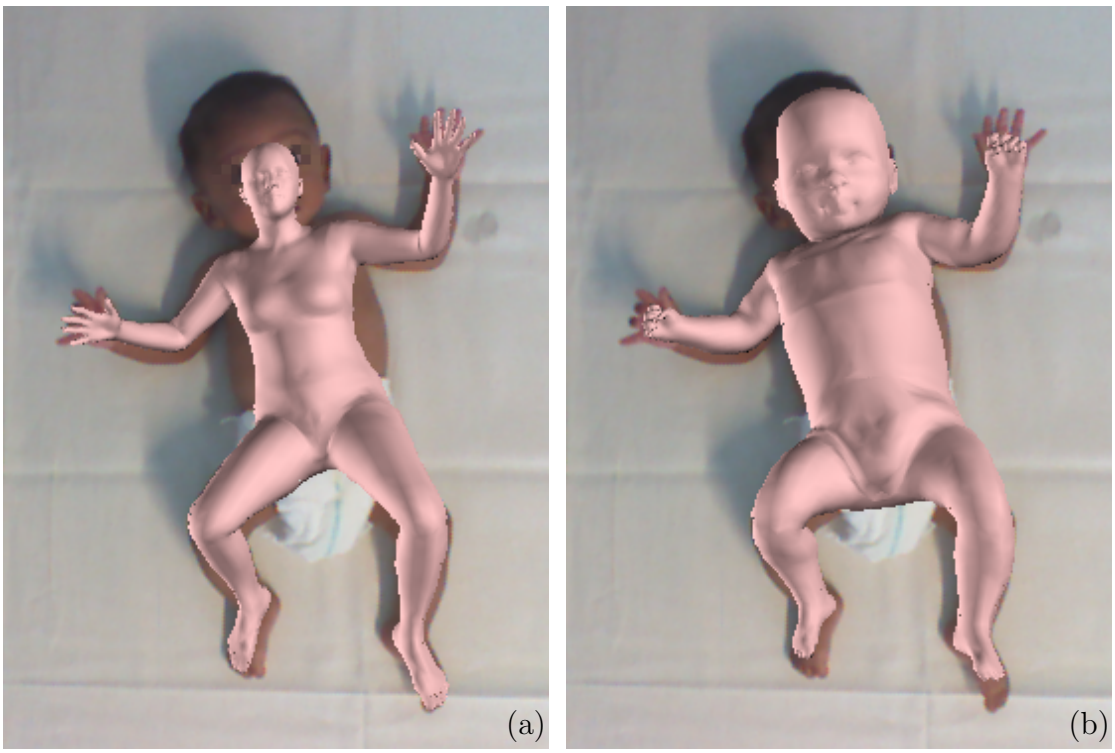


Figure 5.1.: (a) Keep it SMPL [Bog+16]. Result of fitting an adult body model to 2D landmarks of an infant image. (b) Keep it SMIL. Result of fitting our learned infant model to the same image.

5.3. Learning a 3D Infant Body Model from Low Quality RGB-D Data

Parts of this section were published as [Hes+18] and the extended version [Hes+19b].

In the previous section, we have reviewed methods for the creation of human body models, and approaches for using a body model to estimate pose (and shape) of humans from video/depth data.

In order to apply such methods to infant data, we need to create an infant body model. We have discussed the reasons for the non-existence of a data set of high quality infant 3D scans in Sec. 2.4, which is the reason why we can not rely on standard techniques for model creation. Therefore, we present a new method for learning a statistical *Skinned Multi-Infant Linear body model* (SMIL) from RGB-D data of freely moving infants. This means that we want to learn the space of possible shape deformations of infant bodies as well as the range of possible poses.

Learning a body model from data is a chicken-and-egg problem. We need a model to register the data to a common topology, and we need registrations to learn a model. Since no infant body model is available, we first create an initial infant model by

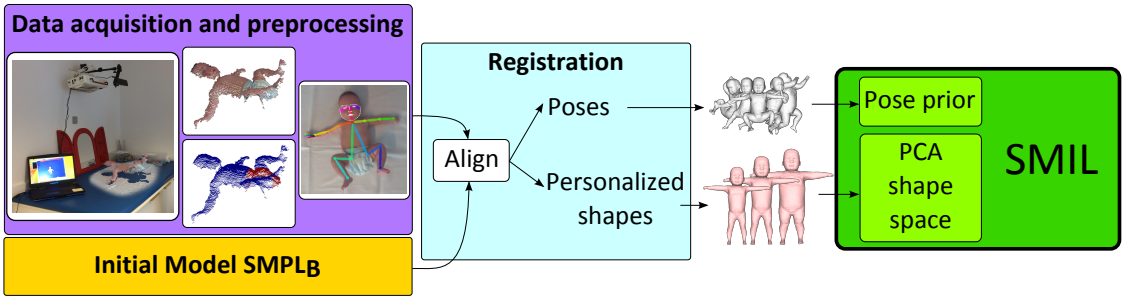


Figure 5.2.: Skinned Multi-Infant Linear model creation pipeline. We create an initial infant model based on SMPL. We perform background and clothing segmentation of the recorded sequences in a preprocessing step, and estimate body, face, and hand landmarks in RGB images. We register the initial model, SMPL$_B$, to the RGB-D data, and create one personalized shape for each sequence, capturing infant shape details outside the SMPL shape space. We learn a new infant specific shape space by performing PCA on all personalized shapes, and a prior on plausible poses from a sampled subset of all poses. Together with the mean of all shapes as base template, this forms the Skinned Multi-Infant Linear model, SMIL.

adapting the adult model SMPL [Lop+15] (see Sec. 5.1.2). We then register this initial model to RGB-D sequences of moving infants (Sec. 5.3.3). To mitigate the incompleteness of data due to the monocular setup, we accumulate shape information from each sequence into one personalized shape (Sec. 5.3.5). Finally, we learn a new infant shape space from all personalized shapes, as well as a new prior on plausible infant poses from our registrations (Sec. 5.3.6). An overview of the complete learning pipeline is given in Fig. 5.2.

5.3.1. Data Acquisition and Preprocessing

There are multiple reasons why no public repository of infant 3D scans exists. Protection of privacy of infants is more strict as compared to adults. The high cost of 3D scanners prevents them from being widespread. Creating a scanning environment that takes into consideration the special care required by infants, like warmth and hygiene, requires additional effort. The rapid growth of infants limits the time frame and the population of possible scanning subjects. Finally, infants can not be instructed to strike poses on demand, which is usually required in standard body scanning protocols. RGB-D sensors offer a cheap and lightweight solution for scanning infants, only requiring the sensor and a connected laptop. The data used to learn SMIL was obtained by setting up recording stations at a children's hospital where infants and parents regularly visit for examinations. The acquisition protocol was integrated in the doctors medical routine in order to minimize overhead. In an ongoing study, we are collecting more RGB-D data by taking advantage of

the lightweight recording setup: we capture the infant's motions at their homes to minimize stress and effort for both infants and parents.

Preprocessing. We transform depth images to 3D point clouds using the camera calibration. To segment the infant from the scene, we fit a plane to the background table of the 3D point cloud using RANSAC [FB81] and remove all points close to or below the table plane and apply a simple cluster-based filtering. Further processing steps operate on this segmented cloud, in which only points belonging to the infant remain. Plane-based segmentation is not always perfect, e.g., in case of a wrinkled towel very close to the infant body some noise may remain. However, the registration methods have shown to be robust to this kind of outliers. The estimated table plane will be reused for constraining the infants' backs in the registration stage (Sec. 5.3.3).

In order to avoid diapers and clothing wrinkles in the infant shape space, we segment the input point clouds into *clothing* and *skin* using the color information by adapting the method from Pons-Moll et al. [PM+17]. We start by registering the initial model to one scan and perform an unsupervised k-means clustering to obtain the dominant modes. We manually define the clothing type to be: naked, diaper, onesie long, onesie short or tights. This determines the number of modes and the cloth prior. The dominant modes allow to create probabilities for each 3D point to be labeled as cloth or skin. We transfer the points' probabilities to the model vertices, and solve a minimization problem on a Markov random field defined by the model topology. We transfer the result of the model vertices to the original point cloud, and we obtain a clean segmentation of the points belonging to clothing (or diaper) and the ones belonging to the skin. An example of the segmentation result can be seen in the *Data acquisition and preprocessing* box of Fig. 5.2. To avoid registering all scans twice, i.e., a first rough registration to segment the sequence and a second to obtain the final registration, we transfer the clothing labels from the registration at frame $t-1$ to the point cloud at frame t. In practice this works well, since changes of the clothed body parts in consecutive frames are relatively small.

The scanned infants can not take a predefined pose to facilitate an initial estimate of model parameters. However, existing approaches on 2D pose estimation from RGB images (for adults) have achieved impressive results. Most interestingly, experiments show that applying these methods to images of infants produces accurate estimates [Hes+19a], cf. Sec. 6.4. In order to choose a "good" candidate frame to initialize the model parameters (see Sec. 5.3.4), we leverage the 2D body landmarks together with their confidence values. From the RGB images we extract body pose [Cao+17] as well as face [Wei+16] and hand [Sim+17] landmarks. We experimentally verify that they provide key information on head and hand rotations to the registration process which is complementary to the noisy point clouds.

5.3.2. Initial Model

We manually create an initial model, which we denote as $SMPL_B$, by adapting SMPL [Lop+15] to infants.

We manually create an infant mesh using makeHuman [Mak] – an open source software for creating 3D characters – which we need to register to SMPL in order to transfer its topology. Directly registering SMPL to the infant mesh fails due to differences in size and proportions. We make use of the fact that meshes exported from makeHuman share the same topology, independent of shape parameters. We register SMPL to an adult makeHuman mesh, and describe makeHuman vertices as linear combinations of SMPL vertices. This allows us to apply this mapping to the infant mesh and transfer it to the SMPL topology. We then replace the SMPL base template geometry with the geometry of the created infant mesh.

We further scale the SMPL pose blend shapes, which correct skinning artifacts and pose-dependent shape deformations, to infant size. We keep the SMPL joint regressor untouched, since it has shown to work well for infants in our experiments. SMPL pose priors, i.e., prior probabilities of plausible poses, are learned from data of adults in upright positions, and therefore can not be directly transferred to lying infants. We manually adjust them based on observations from our experiments. Without adjusting these priors, the model tries to explain shape deformations with pose parameters.

5.3.3. Registration of a Body Model with a 3D Point Cloud

We register the initial model to the segmented point cloud using gradient-based optimization. The main energy being optimized w.r.t. shape β and pose θ parameters is

$$\mathrm{E}(\beta, \theta) = \mathrm{E}_{\mathrm{data}} + \mathrm{E}_{\mathrm{lm}} + \mathrm{E}_{\mathrm{table}} + \mathrm{E}_{\mathrm{sm}} + \mathrm{E}_{\mathrm{sc}} + \mathrm{E}_{\beta} + \mathrm{E}_{\theta}, \tag{5.15}$$

where the weight factors λ_{term}, with term $\in \{\mathrm{data}, \mathrm{lm}, \mathrm{table}, \mathrm{sm}, \mathrm{sc}, \beta, \theta\}$, which are associated with E_{term} are omitted for compactness. In the following, we explain each term of the energy in detail.

Data term. The data term $\mathrm{E}_{\mathrm{data}}$ consists of two different terms:

$$\mathrm{E}_{\mathrm{data}} = \lambda_{\mathrm{m2s}} \mathrm{E}_{\mathrm{m2s}} + \mathrm{E}_{\mathrm{s2m}}. \tag{5.16}$$

$\mathrm{E}_{\mathrm{m2s}}$ accounts for the distance of the visible model vertices to the scan points. $\mathrm{E}_{\mathrm{s2m}}$ accounts for the distance of the scan points to the model surface and

$\mathrm{E}_{\mathrm{m2s}}$ can be written as

$$\mathrm{E}_{\mathrm{m2s}}(M, P) = \sum_{m \in \mathrm{vis(M)}} \rho(\min_{v \in P} ||(m, v)||), \tag{5.17}$$

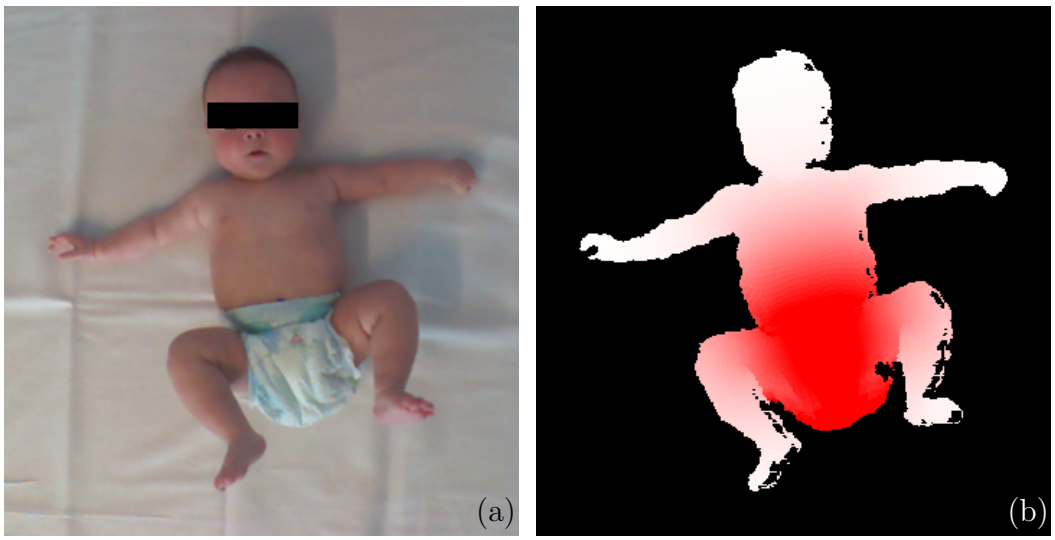


Figure 5.3.: (a) Original RGB image. (b) Weights used for weighted PCA. White points have a high weight (value of 3), red point have a low weight (value of 1). The smooth transition is computed using the skin weights W.

where M denotes the model surface and P the scan points. $||(m, v)||$ is the Euclidean distance between each visible model point and its closest scan point, and ρ is the robust Geman-McClure function [**GemanMcClure1987**]. In the preprocessing stage, P is segmented into the scan points belonging to the skin (P_{skin}) and the ones belonging to clothing (P_{cloth}). The function vis(M) is a function that returns the visible model vertices of M. The visibility is computed using the Kinect V1 camera calibration and the OpenDR renderer [LB14].

E_{s2m} consists of two terms,

$$\text{E}_{\text{s2m}} = \lambda_{\text{skin}}\text{E}_{\text{skin}} + \lambda_{\text{cloth}}\text{E}_{\text{cloth}}. \tag{5.18}$$

E_{skin} enforces the skin points to be close to the model mesh and E_{cloth} enforces the cloth points to be outside the model mesh. The skin term can be written as

$$\text{E}_{\text{skin}}(M, P_{\text{skin}}, W) = \sum_{v \in P_{\text{skin}}} W_v \rho(\text{dist}(v, M)), \tag{5.19}$$

where $\text{dist}(v, M)$ gives the minimal distance of vertex v to the model mesh surface M, and W_v are the skin weights associated with v, which are computed using the method in [Zha+17]. These are used to avoid learning clothing wrinkles in the shape space (cf. Sec. 5.3.6). We assign low weights to clothing points and high weights to skin points with smooth transition weights in between to avoid discontinuities at the clothing borders. Fig. 5.3 displays a sample RGB image and the corresponding skin weights.

The cloth term is divided into two more terms, depending on cloth points lying inside or outside the model mesh:

$$E_{cloth} = E_{outside} + E_{inside}, \tag{5.20}$$

with $E_{outside}$ applying a Geman-McClure penalty to cloth points lying *outside* the body. This is to keep the model from trying to match clothing points that are far away from the body. At the same time, nearby points are closely matched.

$$\mathrm{E}_{\mathrm{outside}}(M, P_{\mathrm{cloth}}, W) = \sum_{v \in P_{\mathrm{cloth}}} \delta_v^{\mathrm{out}} W_v \rho(\mathrm{dist}(v, M)), \tag{5.21}$$

with δ_v^{out} an indicator function, yielding 1 if v lies outside the model mesh, else 0.

The term E_{inside} is used to push the model *inside* the cloth points. We apply a quadratic penalty to cloth points lying *inside* the body mesh, since clothing is not supposed to penetrate the body at all.

$$\mathrm{E}_{\mathrm{inside}}(M, P_{\mathrm{cloth}}, W) = \sum_{v \in P_{\mathrm{cloth}}} \delta_v^{\mathrm{in}} W_v \mathrm{dist}(v, M)^2, \tag{5.22}$$

with δ^{in} an indicator function, returning 1 if v lies inside the model mesh, else 0. Only using E_{inside} would favor shrinking the model too much.

Landmark term. Due to the low quality of data, depth only methods can not reliably capture details like head or hand rotations. However, we can estimate 2D landmark positions from the RGB images and use them as additional constraints in the optimization energy. Body landmarks [Cao+17] are used for initialization (Sec. 5.3.4), whereas face [Wei+16] and hand [Sim+17] landmarks are used in the registration energy of every frame. In the cases where the face detection [Wei+16] fails, mostly profile faces, we use the ears and eyes information from the body pose estimation method [Cao+17]. These help to guide the head rotation in these extreme cases.

The landmark term E_{lm} is similar to Eq. 2 from [Bog+16], where the distances between the 2D landmarks estimated from RGB and the corresponding projections of the 3D model landmarks are measured. Instead of using the body joints, we only use the estimated 2D face landmarks (nose, eyes outlines, mouth outline and ears) as well as the hand landmarks (knuckles). We note the set of all markers as L. The 3D model points corresponding to the above landmarks were manually selected through visual inspection. They are projected into the image domain using the camera calibration matrix in order to compute the final 2D distances to the estimated landmarks.

The landmark term is then

$$\mathrm{E}_{\mathrm{lm}} = \lambda_{\mathrm{lm}} \sum_{l \in \mathrm{L}} c_l \rho(l_M - l_{\mathrm{est}}), \tag{5.23}$$

where c_l denotes the confidence of an estimated landmark 2D location l_{est}, and l_M is the projected model landmark location. All confidences from the different methods are in the interval $[0, 1]$, making them comparable in terms of magnitudes.

Table term. The recorded infants are too young to roll over, which is why the back is rarely seen by the camera. However, the table on which the infants lie, lets us infer shape information of the back. We assume that the body can not be inside the table, and that a large part of the back will be in contact with it. The table energy has two terms: E_{in} prevents the model vertices M from penetrating the table, i.e., behind the estimated table plane, by applying a quadratic error term on points lying inside the table. $\mathrm{E}_{\mathrm{close}}$ acts as a gravity term, by pulling the model vertices M in the close neighborhood of the table towards the table, by applying a robust Geman-McClure penalty function to the model points that are close to the table.

We write the table energy term as

$$\mathrm{E}_{\mathrm{table}} = \lambda_{\mathrm{in}}\mathrm{E}_{\mathrm{in}} + \lambda_{\mathrm{close}}\mathrm{E}_{\mathrm{close}}, \tag{5.24}$$

with

$$\mathrm{E}_{\mathrm{in}}(\mathrm{M}) = \sum_{x \in M} \delta_x^{\mathrm{in}}(x)\mathrm{dist}(x, \Pi)^2, \tag{5.25}$$

and

$$\mathrm{E}_{\mathrm{close}}(\mathrm{M}) = \sum_{x \in M} \delta_x^{\mathrm{close}}(x)\rho(\mathrm{dist}(x, \Pi)), \tag{5.26}$$

where the table plane is denoted as Π. δ_x^{in} is an indicator function, returning 1 if x lies inside the table (behind the estimated table plane), or 0 otherwise and $\delta_x^{\mathrm{close}}$ is an indicator function, returning 1 if x is close to the table (distance less than 3 cm) and faces away from the camera, or 0 otherwise.

To account for soft tissue deformations of the back, which are not modeled, we allow the model to virtually penetrate the table. We effectively enforce this by translating the table plane by 0.5 cm, i.e., pushing the virtual table back.

The weight of the table term needs to be balanced with the data term to avoid a domination of the gravity term, keeping the body in contact with the table while the data term suggests otherwise.

Other terms. Depth data contains noise, especially around the borders. To avoid jitter in the model caused by that noise, we add a temporal pose smoothness term. It avoids important changes in pose unless one of the other terms has strong evidence. The temporal pose smoothness term E_{sm} is the same as in Eq. 21 in [Rom+17] and penalizes large differences between the current pose θ and the pose from the last processed frame θ'. The penalty for model self intersections E_{sc} and the shape prior term E_β are the same as in Eq. 6 and Eq. 7 in [Bog+16] respectively. Bending the model in unnatural ways might decrease the data term error, which is why the pose prior term keeps the pose parameters in a realistic range. The SMIL pose prior consists of mean and covariance that were learned from 37 K sample poses. E_θ penalizes the squared Mahalanobis distance between θ and the pose prior, as described in [Bog+15].

Optimization. To compute the registrations of a sequence, we start by estimating an initial shape using 5 frames. In this first step, we optimize for the shape and pose parameters, β and θ, as well as the global translation t. The average shape parameters from these 5 frames will be kept fixed and used later on as a shape regularizer. Experiments showed that otherwise the shape excessively deforms in order to explain occlusions in the optimization process.

With the initial shape fixed, we compute the poses for all frames in the sequence, i.e., we optimize the following energy w.r.t. pose parameters θ and the global translation t:

$$\mathrm{E}(\theta, \mathrm{t}) = \mathrm{E}_{\mathrm{data}} + \mathrm{E}_{\mathrm{lm}} + \mathrm{E}_{\mathrm{table}} + \mathrm{E}_{\mathrm{sm}} + \mathrm{E}_{\mathrm{sc}} + \mathrm{E}_{\theta}. \tag{5.27}$$

Notice that this energy is equal to Eq. 5.15 without the shape term $\mathrm{E}_{\mathrm{beta}}$, as shape is kept fixed. We denote S_f the computed posed shape at frame f.

In the last step, we compute the registration meshes R_f and allow the model vertices $v \in R_f$ to freely deform to best explain the input data. We optimize w.r.t. v the energy

$$\mathrm{E}(v) = \mathrm{E}_{\mathrm{data}} + \mathrm{E}_{\mathrm{lm}} + \mathrm{E}_{\mathrm{table}} + \mathrm{E}_{\mathrm{cpl}}, \tag{5.28}$$

where $\mathrm{E}_{\mathrm{cpl}}$ is used to keep the registration edges close to the edges of the initial shape. We use a similar energy term as Eq. 8 in [Bog+15]:

$$\mathrm{E}_{\mathrm{cpl}}(R_f, S_f) = \lambda_{cpl} \sum_{e \in V'} ||(AR)_e - (AS)_e||_F^2, \tag{5.29}$$

where V' denotes the edges in the model mesh. AR and AS are edge vectors of the triangles of R_f and S_f, and e indexes the edges. The results of these optimizations are the final registrations.

All energies are minimized using a gradient-based dogleg minimization method [NW06] with OpenDR [LB14] and Chumpy [Lop]. We display registration samples in Fig. 5.4.

5.3.4. Initializing the Registration Process

In order to find the global minimum, the optimization needs a good initial estimate. In adult settings, subjects are usually asked to take an easy pose, e.g., T-pose (extended arms and legs), at the start of the recording. Infants are not able to strike poses on demand, which is why we can not rely on a predefined pose.

We automatically find an initialization frame containing an "easy" pose by relying on 2D landmark estimates acquired in the preprocessing stage. We make the assumption that a body segment is most visible if it has maximum 2D length over the

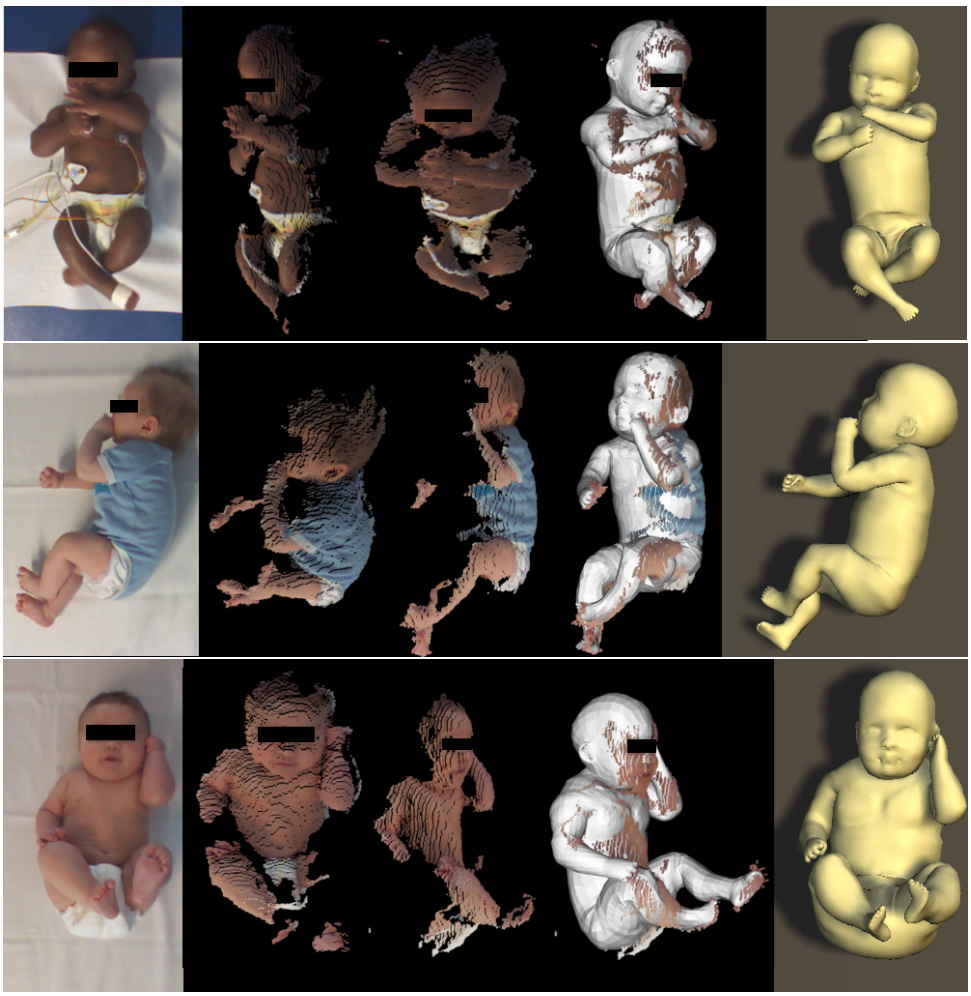


Figure 5.4.: SMIL Registration result samples. From left to right: RGB, point cloud, point cloud (other view), point cloud with registered SMIL, rendered registration.

complete sequence. Perspective projection would decrease 2D body segment length and therefore visibility. The initialization frame is chosen as

$$f_{\text{init}} = \text{argmax}_f \sum_{s \in S} \text{len}(s, f) c(s, f), \tag{5.30}$$

where S is the set of body segments, $\text{len}(s, f)$ is the 2D length of the segment s in frame f, and $c(s, f)$ is the estimated confidence of joints belonging to s in frame f.

For f_{init}, we optimize a simplified version of Eq. 5.15, i.e., the initialization energy

$$\text{E}_{\text{init}} = \lambda_{\text{j2d}}\text{E}_{\text{j2d}} + \lambda_\theta \text{E}_\theta + \lambda_{\text{a}}\text{E}_{\text{a}} + \lambda_\beta \text{E}_\beta + \lambda_{\text{s2m}}\text{E}_{\text{s2m}} \tag{5.31}$$

where E_{j2d} is similar to E_{lm} with landmarks being 2D body joint positions. E_θ is a pose prior with high weight λ_θ, $\text{E}_{\text{a}}(\theta) = \sum_i \exp(\theta_i)$ is an angle limit term for knees

and elbows and E_β is a shape prior. Its minimum provides a coarse estimation of shape and pose which is refined afterwards. In contrast to [Bog+16], we omit the self intersection term, and add a scan-to-mesh distance term E_{s2m}, containing 3D information, while [Bog+16] solely relies on 2D information.

5.3.5. Creation of Personalized Shapes

To capture the subject specific shape details, we create one personalized shape from each sequence, which we do not restrict to the shape space of the model. We *unpose* a randomly selected subset of 1000 frames per sequence. The process of unposing changes the model pose to a normalized pose (T-pose) in order to remove variance related to body articulation. For each scan point, we calculate the offset normal to the closest model point. After unposing the model, we add these offsets to create the unposed point cloud for each of the 1000 frames. Since the recorded infants lie on their backs most of the time, the unposed clouds have missing areas on the back side. To take advantage of the table constraint in each frame and sparsly fill the missing areas, we add *virtual points*, i.e., points from model vertices that belong to faces oriented away from the camera, to the unposed cloud. We retain the clothing segmentation labels for all unposed scan points. We call the union of all unposed point clouds including virtual points the *fusion cloud*.

To compute the personalized shape, we uniformly random sample one million points from the fusion cloud and proceed in two stages. First, we optimize $E = E_{data} + E_\beta$ w.r.t. the shape parameters β, and keep the pose θ fixed in the zero pose of the model. We obtain an initial shape estimate that lies in the shape space of the initial model $SMPL_B$. Second, we allow the model vertices to deviate from the shape space, but tie them to the shape from the first stage with a coupling term. We optimize $E = E_{data} + E_{cpl}$ w.r.t. the vertices.

The clothing segmentation is also transformed to the unposed cloud and therefore, the fusion cloud is labeled into clothing and skin parts. These are used in the data term to enforce the clothing points to lie ouside the model surface and to avoid learning clothing artifacts in the shape space.

5.3.6. Learning the Skinned Multi-Infant Linear Model Shape Space and Pose Prior

We compute the new infant shape space by doing weighted principal component analysis on personalized shapes of all sequences. Despite including the clothing segmentation in the creation of personalized shapes, clothing deformations can not be completely removed and diapers typically tend to produce body shapes with an over-long trunk. The recorded sequences contain infants wearing long-arm onesies, short-arm onesies, tights, diapers and infants without clothing. These different

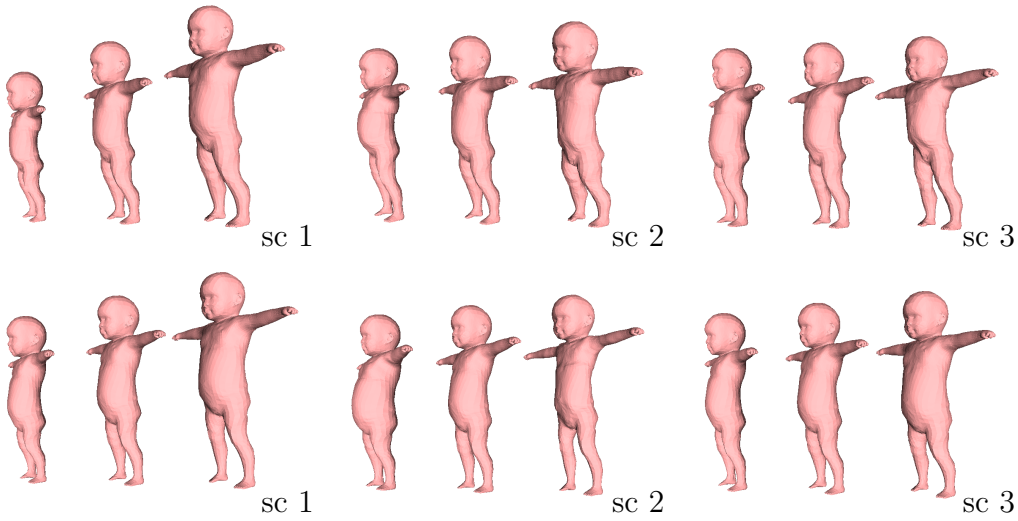


Figure 5.5.: First three shape principal components (sc). Top: SMIL, -2 to +2 standard deviations. Bottom: SMPL_B, -0.5 to +0.5 standard deviations. The first components in the infant shape (SMIL sc 2 and 3) carry variation in trunk size/length, while the first components of SMPL_B show trunk variation mainly in the belly growing or shrinking.

clothing types cover different parts of the body. As we want the shape space to be close to the real infant shape without clothing artifacts, we use low weights for clothing points and high weights for skin points in the PCA. The weights we use to train the model are: 3 for the scan points labeled as skin (P_skin), 1 for the scan points labeled as clothing (P_cloth), and we compute smooth transition weights for the scan points near the cloth boundaries using the skin weights W computed using the method in [Zha+17]. Fig. 5.3 displays the weights used for the weighted PCA on a sample frame. We use the EMPCA algorithm[1] computing weighted PCA with an iterative expectation-maximization approach. We retain the first 20 shape components. We display the first 3 shape components for SMIL and for SMPL_B in Fig. 5.5.

We create a pose data set by looping over all poses of all sequences and only add poses to the set if the dissimilarity to any pose in the set is larger than a threshold. The final set contains 47K poses and is used to learn the new pose prior. As the pose prior can not extrapolate and penalize illegal poses, e.g., unnatural bending of knees, we manually add penalties to avoid such poses.

The final SMIL model is composed of the shape space, the pose prior, and the base template, which is the mean of all personalized shapes.

[1] `https://github.com/jakevdp/wpca`

5.3.7. Manual Intervention

In our method we use manual intervention three times: i) to decide which type of clothing the infant is wearing (see Sec. 5.3.1); ii) to generate the initial model SMPL_B (see Sec. 5.3.2) and iii) to define illegal poses in the pose prior. The illegal poses are only defined once and the initial model is no longer used once SMIL is learned. However, given a new sequence, one still needs to manually define the type of clothing: short onesie, long onesie, tights, naked or diapers. Each cloth type defines the corresponding number of color modes and priors to be used. While this is the only remaining manual step in our method, we believe that a classifier predicting the clothing type from RGB images could be learned, making our method fully automatic.

5.3.8. Registration Objective Function Weights

The values of the weights in the energy functions were empirically adjusted to keep the different terms balanced.

For optimization of the main energy w.r.t. shape and pose parameters (Eq. 5.15) and the modified energy w.r.t. pose parameters (Eq. 5.27) we use the weight values: $\lambda_\text{skin} = 800$, $\lambda_\text{cloth} = 300$, $\lambda_\text{m2s} = 400$, $\lambda_\text{lm} = 1$, $\lambda_\text{table} = 10000$, $\lambda_\text{sm} = 800$, $\lambda_\text{sc} = 1$, $\lambda_\beta = 1$ and $\lambda_\theta = 0.15$.

For optimization of the energy w.r.t. the model vertices (Eq. 5.28) we use the weight values: $\lambda_\text{skin} = 1000$, $\lambda_\text{cloth} = 500$, $\lambda_\text{m2s} = 1000$, $\lambda_\text{lm} = 0.03$, $\lambda_\text{table} = 10000$ and $\lambda_\text{cpl} = 1$.

For the creation of the personalized shape (Sec. 5.3.5), we use weight values: $\lambda_\text{skin} = 100$, $\lambda_\text{cloth} = 100$ $\lambda_\beta = 0.5$ and $\lambda_\text{cpl} = 0.4$.

Finally, for the initialization energy (Eq. 5.31), we use: $\lambda_\text{j2d} = 6$, $\lambda_\theta = 10$, $\lambda_\text{a} = 30$, $\lambda_\beta = 1000$, $\lambda_\text{s2m} = 30000$.

We keep the chosen weights constant for all experiments.

5.4. Evaluation

As elaborated in the introduction, gathering high quality 3D scans of infants is highly unpractical, which is why we quantitatively evaluate SMIL and our initial model SMPL_B on the 37 acquired RGB-D sequences of infants. We record the infants using a Microsoft Kinect V1, which is mounted 1 meter above an examination table, facing downwards. All parents gave written informed consent in participating in this study, which was approved by the ethics committee of Ludwig Maximilian University Munich (LMU). The infants lie in supine position for three to five minutes without external stimulation, i.e., there is no interaction with caregivers or toys. The

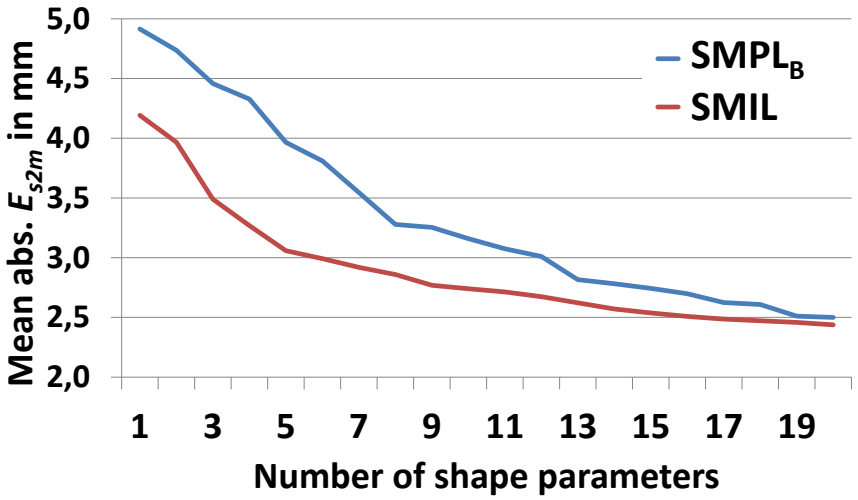


Figure 5.6.: Average scan-to-mesh error Es2m in mm w.r.t. the number of shape parameters for the two models registered to all fusion scans.

recorded infants are between 9 and 18 weeks of corrected age (post term), and their size range is 42 to 59 cm, with an average of 53.5 cm. They wear different types of clothing: none, diaper, onesie shortarm/longarm, or tights. All sequences together sum up to roughly 200K frames, and have an overall duration of over two hours. We evaluate SMIL with a 9-fold cross-validation, using 33 sequences for training and 4 for testing and we distribute different clothing styles across all sets. We measure the distance E_{s2m} (cf. Eq. 5.18) of the scan to the model mesh by computing the Euclidean distance of each scan point to the mesh surface. For evaluation, we consider all scan points to be labeled as skin, which reduces Eq. 5.18 to Eq. 5.19. Note that we do not use the Geman-McClure function ρ here, as we are interested in the actual Euclidean distances.

To compare the SMPL$_B$[2] shape space to the SMIL shape space, we register both models to each of the 37 *fusion scans*, using different numbers of shape components. Results are displayed in Fig. 5.6. We plot average error heatmaps for using the first 1, 3, 5 and all 20 shape components for the registrations (Fig. 5.7). We observe lower error for SMIL for smaller numbers of shape parameters, and a nearly identical error when using all 20 parameters. The small difference between SMPL$_B$ and SMIL poses the question if there is a significant difference between the two at all. A visual comparison of the registration with all 20 shape components shows that SMPL$_B$ shapes display more adult-like features, like a more pronounced waisting, a more "muscular" chest and shoulder region, as well as a smaller (fore)head area. We show some examples in Fig. 5.8.

[2]Note: SMPL$_B$ is not the SMPL model [Lop+15], but our initial infant model, registered to the data using our method.

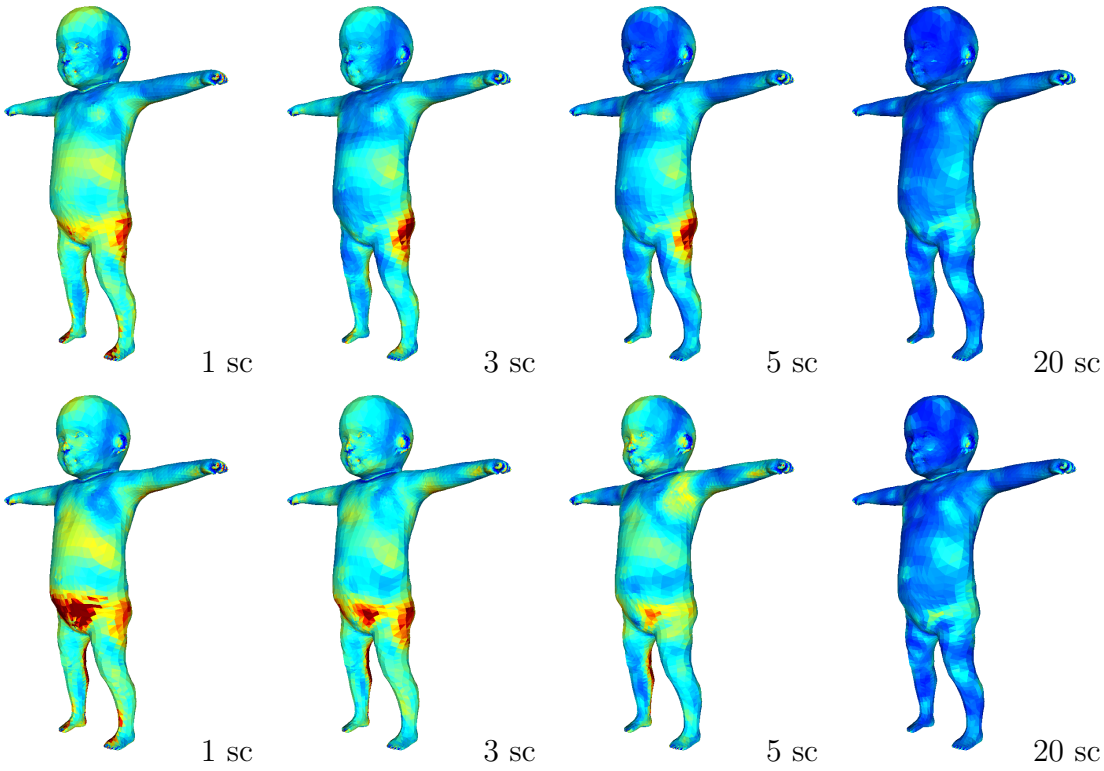


Figure 5.7.: Average error heatmaps for SMIL and $SMPL_B$ on fusion clouds for different numbers of shape components (sc). Top: SMIL. Bottom: $SMPL_B$. Blue means 0 mm, red means $\geq$ 10 mm.

To evaluate how well the computed personalized shapes and poses explain the input sequences, we calculate E_{s2m} for all 200K frames. SMIL achieves an average scan-to-mesh distance of 2.51 mm (SD 0.21 mm), $SMPL_B$ has an average E_{s2m} of 2.67 mm (SD 0.22 mm). This is in the range of Bogo et al. [Bog+15], who report an average error between 2.45mm (static scans) and 3.23 mm (arbitrary motions) on adults, captured with a Kinect V2.

Due to the lack of ground truth data for evaluation of infant pose correctness, we perform a manual inspection of all sequences to reveal pose errors. We distinguish between "unnatural poses" and "failure cases". Unnatural poses contain errors in pose, like implausible rotations of a leg (cf. Fig. 5.9 top row), while the overall registration is plausible, i.e., the 3D joint positions are still at roughly the correct position. Failure cases denote situations in which the optimization gets stuck in a local minimum with a clearly wrong pose, i.e., one model body part registered to a scan part which it does not belong to (cf. Fig. 5.9 bottom row). We count 16 unnatural leg/foot rotations lasting 41 seconds (= 0.54% of roughly 2 hours) and 18 failure cases (in 7 sequences) lasting 49 seconds (= 0.66% of roughly 2 hours).

To evaluate how well SMIL generalizes to older infants, we register the model to 25 sequences of infants at the age between 21 and 36 weeks, at an average of 26

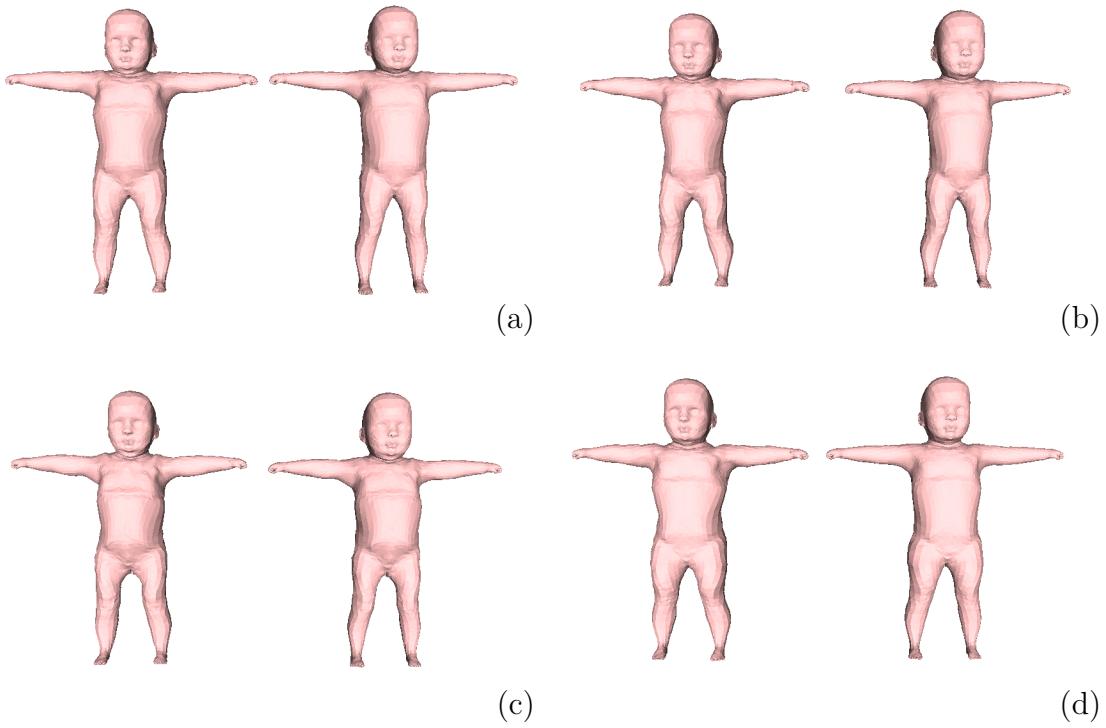


Figure 5.8.: Visual comparison of four $SMPL_B$ (left sides) and SMIL (right sides) registration samples with all 20 shape components. Most visually significant differences occur at waist, chest, shoulders and head.

weeks. The resulting average scan to mesh distance is 2.83 mm (SD: 0.31 mm). With increasing age, infants learn to perform directed movements, like touching their hands, face, or feet, as displayed in Fig. 5.10. This makes motion capture even more challenging, as standard marker-based methods would not be recommended because of the risk of infants grabbing (and possibly swallowing) markers.

Failure cases. The most common failure is a mixup of feet, i.e., left foot of the model registered to the right foot of the scan and vice versa. Despite our energy having the interpenetration penalty E_{sc}, we observe a few cases where the legs interpenetrate, as in the bottom row in Fig. 5.9. The registration of all sequences is time consuming (between 10 and 30 seconds per frame), so rerunning the full 200K registrations many times to optimize the parameters is not feasible. The energy term weights are manually selected in order to balance the different terms, and by visually inspecting the results of some sequences. Further manual adjustment of the E_{sc} weight could fix these rare cases. In the example in the top row of Fig. 5.9, the right knee is twisted in an unnatural way after the right foot was completely occluded. When the foot is visible again, the pose recovers (knee twisted for 5-6 seconds). Similar to the first failure case, a higher weight on the pose prior would prevent such cases, but finding the perfect weight which completely forbids all illegal

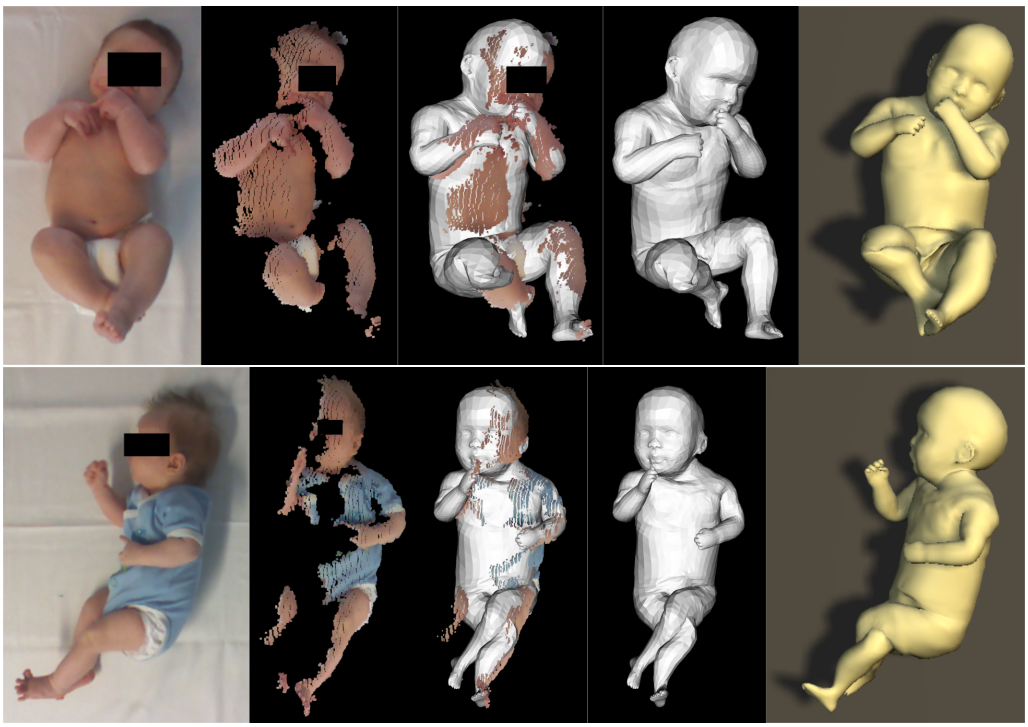


Figure 5.9.: Pose errors. Top: registration result takes an unnatural pose. Bottom: Failure case sample. Right model leg is registered to left scan leg, and vice versa. From left to right: RGB image, 3D point cloud (rotated for improved visibility), overlay of point cloud and registration result, registration result, rendered result from same viewpoint as RGB image.

poses while allowing all legal poses would require a significant engineering effort.

There is no definition of how much detail or accuracy in capturing motions is necessary for medical motion analysis. We argue that SMIL captures sufficient detail if expert's GM ratings of original RGB videos are the same as on rendered SMIL alignment sequences of the same recording.

5.5. General Movement Assessment Case Study

We conduct a case study on GMA to show that SMIL captures enough information to allow medical assessment. Three trained and certified GMA-experts perform GMA in different videos. We use five stimuli: i) the original RGB videos (denoted by V_{rgb}, cf. Fig. 5.4 left), and ii) the synthetic alignment videos (V_{reg}, cf. Fig. 5.4 right). For the next three stimuli we use the acquired poses of infants, but we animate a body with a different shape, namely iii) a randomly selected shape of another infant (V_{other}), iv) an extreme shape producing a very thick and large baby (V_{large}), and v) the mean shape (V_{mean}). We exclude three of the 37 sequences,

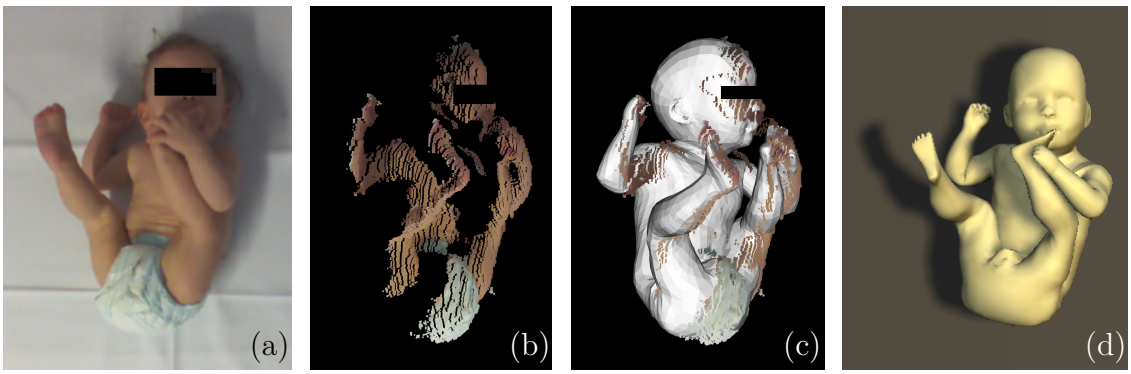


Figure 5.10.: Older infant in very challenging pose. (a) RGB image, (b) 3D point cloud (rotated for improved visibility), (c) overlay of point cloud and SMIL registration result, (d) rendered SMIL registration result.

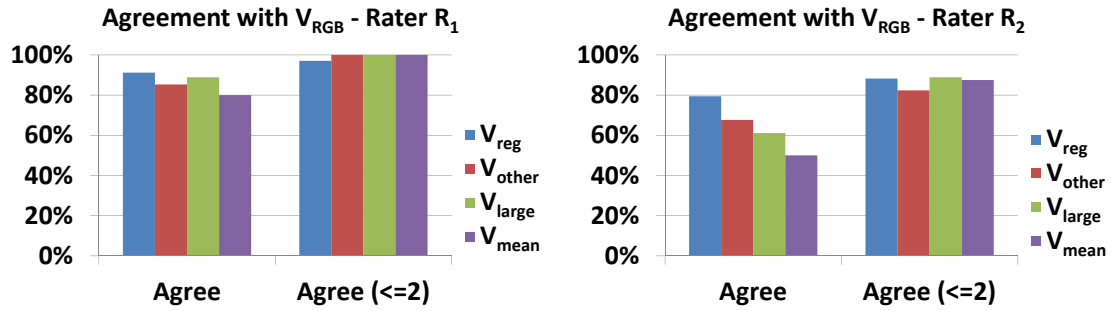


Figure 5.11.: Results of GMA case study. Percentage of ratings of synthetic sequences, generated using SMIL, that *agree* with the reference ratings R_1V_{rgb} (left) and R_2V_{rgb} (right), respectively. $V_{\{reg,other,large,mean\}}$ denotes different stimuli.

as two are too short and one has non-nutritive sucking, making it non suitable for GMA. As the number of videos to rate is high (34*5), for iv) and v) we only use 50% of the sequences, resulting in 136 videos. To avoid recognition of infants from RGB videos leading to any scoring bias, we let experts first rate all randomly shuffled synthetic videos, and then all original RGB videos. Instead of having the experts classify motion quality into one of standard GMA classes *definitely abnormal* (DA), *mildly abnormal* (MA), *normal suboptimal* (NS), and *normal optimal* (NO) [HA04], we augment these four classes with a one to ten scale in order to make a more fine grained evaluation possible. Scores 1-3 correspond to DA, 4-5 to MA, 6-7 to NS, and 8-10 to NO. We consider two ratings with an absolute difference ≤ 1 to *agree*, and otherwise to *disagree*.

Rater R_1 is a long-time GMA teacher and has worked on GMA for over 25 years, R_2 has 15 years experience in GMA, and R_3 was certified one year ago, but lacks clinical routine in GMA. Average of GMA ratings (and standard deviation) for R1 is 4.7 (1.4), for R2 4.0 (1.9), and for R3 4.9 (2.3). The agreement on original RGB ratings V_{rgb} between R_3 and the more experienced raters is lower than 50%, while R_1 and

R_2 agree on 65% of the ratings. R_3 is the treating physician of the infants in the data set. Therefore, it is very probable that he is biased towards the medical history of infants since he recognizes them from the RGB videos. The synthetic videos could to overcome this bias, but results do not show a significant agreement between R_3's and the other expert's ratings of synthetic sequences. We (and R_3 himself) explain the poor results with R_3's lack of experience and regular training in GMA. The low agreement further stresses that GMA is challenging and its automation important. Due to the high rater variability we further focus on ratings of experienced raters R_1 and R_2. In Fig. 5.11, we present rating differences between synthetic and reference sequences. Each rater is compared to her own V_{rgb} ratings as a reference. R_1V_{reg} ratings *agree* on 91% of the reference ratings, whereas R_2 achieves an agreement rate of 79%. The agreement decreases more (R_2) or less (R_1) when the motions are presented with a different body shape. By extending the agreement threshold to ≤ 2, the percentages of all sequences become very similar.

In our case-study, we observe that R_1's and R_2's ratings only agree on $\approx 65\%$ of the original RGB videos V_{rgb} although both are very experienced. In Fig. 5.12 we present the histogram of signed differences between the ratings of all raters. In the first row, we show the differences between R_1 and R_2. We can see that the peak of what looks like a normal distribution is centered at one instead of zero. Our interpretation is that R_2 has the same tendency as R_1, but R_2's ratings are shifted by 1. The comparison of the more experienced raters with R_3 shows a more diffuse distribution of differences, indicating that R_3's ratings are more inconsistent.

We further analyze the signed distances between the ratings of the original RGB videos V_{rgb} and the different synthetic sequences of R_2. Interestingly, we observe that V_{reg}, V_{other} and V_{mean} have in general healthier ratings, whereas V_{large} have less healthy ratings, as shown in Fig. 5.14.

5.6. Discussion

Why does it work? Even though each RGB-D frame is a partial observation of the shape, the motions in the different frames of a sequence reveal previously hidden body parts. Moreover, even though the backs of the infants are rarely visible, we can still faithfully infer where the infants' backs are by taking into account the background table constrains. This lets us accumulate shape information over sequences. In addition, we leverage 2D pose estimation from RGB in two ways: i) it allows us to add landmark constraints to guide the model where depth only approaches fail and ii) allows us to find a good initialization frame and circumvent the need of predefined poses.

Why has it not been done before? Most previous work has been done on adults, who can be instructed and 3D scanned much more easily. Recording infants poses more challenges (see Sec. 5.3.1). As our setup only relies on a low-cost RGB-D camera and

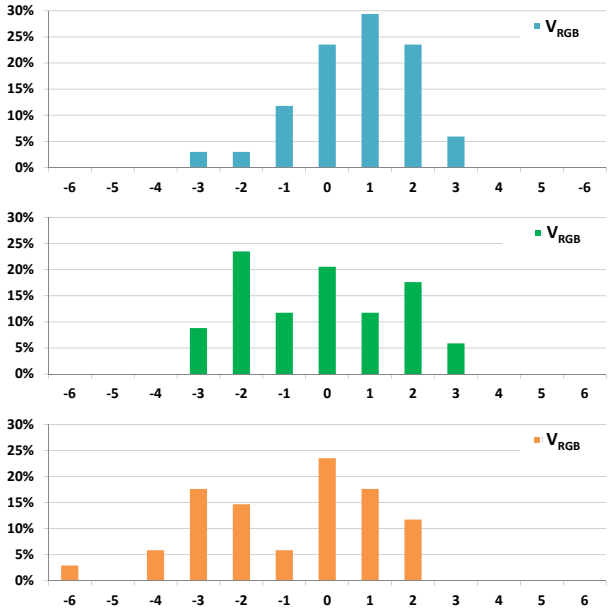


Figure 5.12.: Histogram of the signed differences between R_2 and R_3 for the original RGB videos V_{rgb}.

a laptop, we can capture infants with a great flexibility, for instance at a children's hospital that parents and children are visiting anyway. The creation of an initial model was not straightforward. Thanks to the flexibility of the SMPL model, we were able to adapt it to infants and get a starting point for capturing real infant shapes.

Conclusion. We contribute a method for learning a body model from RGB-D data of freely moving infants. We show that our learned Skinned Multi-Infant Linear model factorizes pose and shape and achieves metric accuracy of 2.5 mm. We further applied the model to the task of medical infant motion analysis. Two expert GMA raters achieve a scoring agreement of 91%, respectively 79%, when comparing the assessment of movement quality from standard RGB video and from rendered SMIL registration results of the same sequence. Our method is a step towards a fully automated system for GMA, to help early-detect neurodevelopmental disorders like cerebral palsy.

Limitations and Future Work. The model can neither produce single finger/toe movements, nor facial expressions. We plan to add this functionality, in a way similar to [Rom+17] and [Li+17]. The question will be whether or not the adult finger movements and adult facial expressions can be simply transferred to the infant model, or if we will have to learn them from infant data.

In our registration results, we found few cases of unusual neck twists. The SMIL

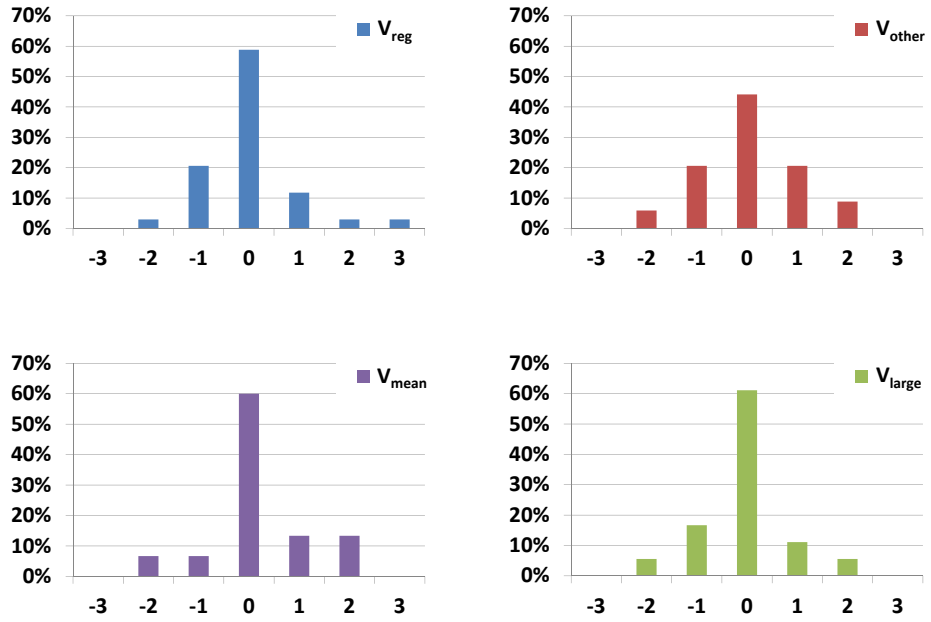


Figure 5.13.: Histograms of signed differences between R_1's ratings of V_{rgb} and R_1's ratings of V_{reg}, V_{other}, V_{mean} and V_{large}. The differences are clearly centered around zero and show no significant bias.

neck seems to be longer than the average infant neck, which is why it is sometimes twisted to achieve a compression to match the pose in the data. These cases are mostly when the infant rotates the head to the side, which in general differs from the (automatically selected) initialization poses, which are mostly frontal. Since we calculate the shape for each sequence on the first 5 frames starting from the initialization frame, it may have not picked up enough information for side views. The model then tries to explain the head shape in the scan, which possibly deviates from the calculated shape, by adjusting the neck pose parameters.

Infant clothing, especially diapers, largely deviates from the body, and SMIL does not explicitly model clothing. When fitting the model to clothed infant data, it deforms to explain the clothing. Our clothing term dampens the effect, but can not completely prevent it. It may be necessary to learn a model of clothing, possibly by recording the same infants with and without clothing, in order to infer the real body shape from the clothed shape.

SMIL was learned from data of infants up to 4 months of age, since this is the relevant age for GMA. From our experiments, we observe that the infant model seems to generalize to older infants, but at some point, the difference in body proportions will become too apparent. Nonetheless, the proposed model and methods can be used to learn models of older infants step by step, to finally close the gap between the infant model and existing adult models.

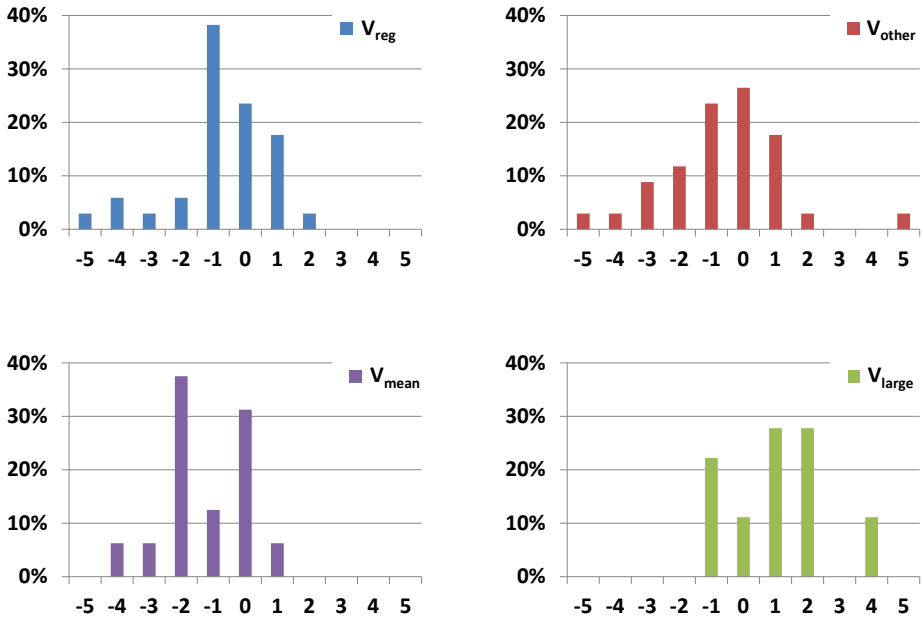


Figure 5.14.: Histograms of signed differences between R_2's ratings of V_{rgb} and R_2's ratings of V_{reg}, V_{other}, V_{mean} and V_{large}. Ratings for V_{large} are less "healthy", while ratings for the other shapes are more "healthy". The different shapes seem to influence R2's ratings.

A different issue with older infants is that they start to turn into prone position. When they turn their back towards the camera, arms and hands are often used to support the body, and are therefore hidden below the body. This makes them invisible to the camera and therefore impossible to track. One could try to guess the position/configuration from prior knowledge about plausible positions. However, our focus in this thesis lies on the general movement assessment, which is only valid for analyzing *spontaneous*, i.e., non-directed, movements. These are generally shown until the age of roughly four months, at which infants usually are not able to turn yet.

We believe that our methods are applicable for tracking older children, who are turning and/or crawling, with an adapted version of the model. In that case, a change of setup, i.e., a change of camera position or the use of multiple cameras, may be necessary to resolve complete and enduring occlusions of limbs.

Our plane-based segmentation method assumes a constrained setup, with a single infant lying on a flat surface, which generally can be easily constructed in doctor's offices. If the scene becomes more cluttered, or a person interacts with the infant, the simple segmentation approach may not be sufficient and one may have to rely on more sophisticated methods for segmentation, e.g., [He+17]. The combination of RGB and depth information should simplify the task.

The method does not run in real-time. Some applications, like monitoring patients for seizure detection, require processing the incoming data in real-time in order to raise an alarm in case of emergency. For motion analysis, the real-time constraint is not essential. Although a low run-time is always desirable, we put the focus on quality of registration and not on speed.

Although failure cases are very rate, they still exist, and may harm the automated prediction of GMA scores. Since one can not guarantee these failures to be completely eliminated, we will strive to detect and preferably correct them, e.g., like [Ari+18a].

In this work we have not learned pose-dependent shape deformations for infants, and we reused the scaled down pose blend shapes of SMPL. While these provide sufficient numerical accuracy, we will learn infant pose blend shapes to further increase the realism of SMIL.

We have not used deep learning for estimating pose (and shape), although it revolutionized many areas of computer vision and consistently improves state of the art results. However, deep neural networks require large sets of training data. With the creation of SMIL it has become possible to create large sets of synthetic data from real poses and real shapes. It is possible to animate each shape with all captured poses to increase the amount of data, or to sample from the shape space, or to interpolate poses, etc. This will allow us to train deep neural networks with realistic infant data in the future.

In the next chapter, we describe the creation of a realistic data set using SMIL for the evaluation of infant motion tracking.

6. Moving INfants In RGB-D (MINI-RGBD) Data Set

In Sec. 2.4, we identified a lack of evaluation in the literature on infant motion capture methods, which we trace back to the non-existence of public infant data sets. The SMIL model described in the previous chapter allows us to create realistic RGB-D data with accurate ground truth joint positions. In this chapter, we describe the methods to generate the data set and provide baseline experiments. This method could be applied at a larger scale to generate enough training data for deep learning approaches.

Parts of this chapter were published as [Hes+19a].

An RGB-D data set for the evaluation of infant pose estimation approaches needs to fulfill several requirements. It has to cover (i) realistic infant movements, (ii) realistic texture, (iii) realistic shapes, and (iv) precise ground truth, while (v) not violating privacy. Our presented data set fulfills all of these requirements.

The data set creation procedure can be divided into two stages (see Fig. 6.1). First, we capture the shape and pose of infants using SMIL, and map the captured real infant movements to newly generated realistic textured infant shapes. Second, we render realistic RGB and depth images from the created movement sequences, and generate accurate 2D and 3D ground truth joint positions (Sec. 6.2). Two samples are displayed in Fig. 6.2, and an overview of all sequences can be found in Appendix B.

The capturing of shape and pose is achieved by registering SMIL to 12 RGB-D sequences of moving infants.

6.1. Extracting the Body Texture from RGB-D Sequences

We follow the protocol from Sec. 5.3.3 to register SMIL to point clouds created from RGB-D sequences, which we also denote as *scans*.

We extend the previous method by generating one texture for each sequence, similar to [Bog+15] and [Bog+14]. We create a texture map by finding closest points from textured point cloud and registered model mesh, as well as a corresponding normal map for each frame. We merge 300 randomly selected texture maps from each

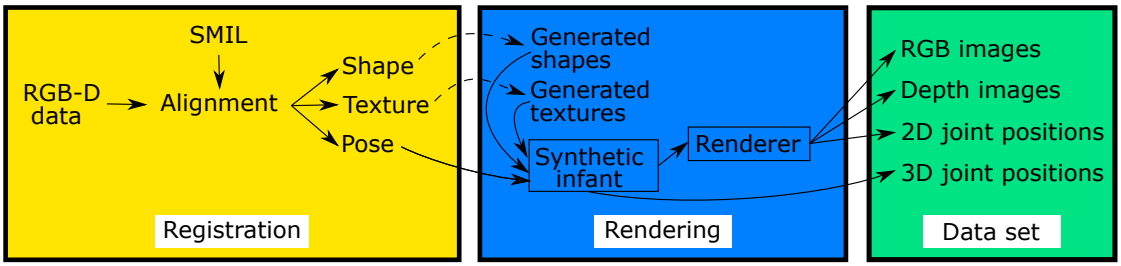


Figure 6.1.: Overview of data set creation pipeline. SMIL body model [Hes+18] is aligned to real infant RGB-D data. Subsets of shapes and textures are used for generating realistic, privacy preserving infant bodies. We animate the new "synthetic infants" with real movements (poses) from the registration stage. We render RGB and depth images, and create ground truth 2D and 3D joint positions to complete our new data set.

sequence by averaging texture maps that are weighted according to their normal maps, with higher weights for points with normals directed towards the camera.

Infants tend to lie on their backs without turning, which is why the merged texture maps have blank areas depending on the amount of movement in the sequence. We fill the missing areas by expanding the borders of existing body areas. To preserve the privacy of the infants we do not use textures from single sequences, but generate average textures from subsets of all textures. The resulting texture maps (sample displayed in Fig. 6.3 (a)) are manually post-processed by smoothing borders and visually enhancing areas of the texture for which the filling did not create satisfying results. We also create average shapes from subsets of shapes from the registration stage to create a variety of realistic body shapes, as shown in Fig. 6.3 (b).

6.2. Rendering RGB and Depth Images with Ground Truth Body Joint Positions

For each of the 12 sequences, we randomly select one of the average shapes and one of the average textures. We map the pose parameters of the sequence, obtained in the registration stage, to the new shape, and animate textured 3D meshes of realistic, yet artificial infants. Based on plane parameters extracted from the background table of the real sequences, we add a plane to simulate the background. We texture the plane with one of various background images (e.g., examination table, crib, changing table) to account for background variation. We use OpenDR [LB14] to render RGB and depth images from the meshes and backgrounds. We select 1000 consecutive frames from each sequence where the infant is the most active. The rendered depth image is overly smooth, which is why we add random noise of up to ± 0.5 cm to simulate noise levels of depth sensors. We use camera parameters similar

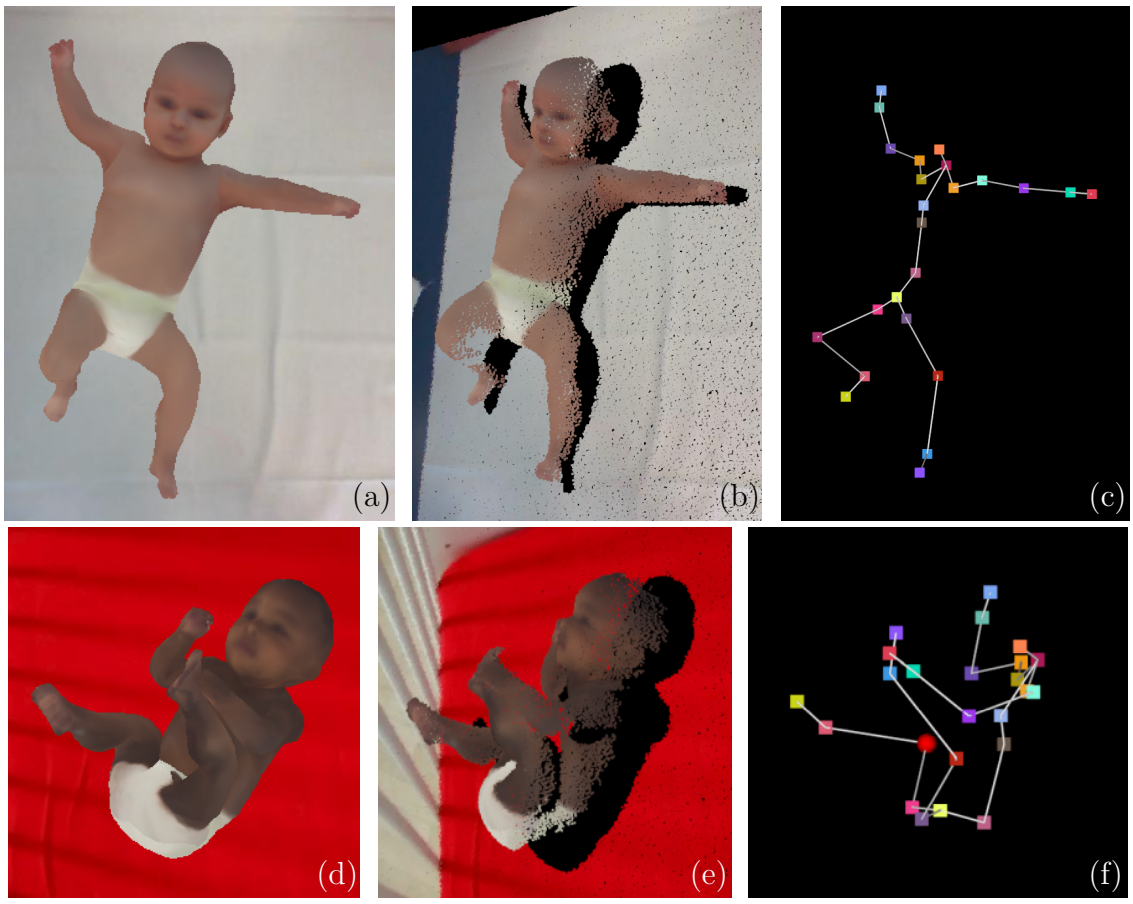


Figure 6.2.: Two samples from MINI-RGBD data set. (a) and (d): RGB image. (b) and (e): point cloud created from depth image. (c) and (f): ground truth skeleton. Viewpoint for (b), (c), (e), and (f) is slightly rotated to side.

to Microsoft Kinect V1, which is the most frequently used sensor in approaches in Sec. 2.3.3, at a resolution of 640 * 480, at 30 frames per second. The distance of the table to the camera is roughly 90 cm for all sequences. The 3D joint positions are directly regressed from the model vertices (see Fig. 6.2 (c) and (f)). To provide 2D ground truth, we project the 3D joints to 2D using the camera calibration. For completeness, we add depth values for each joint. To simplify data set usage, we provide a segmentation mask discriminating between foreground and background.

6.3. Data Set Summary

We generate 12 sequences, each with different shape, texture, movements, and background, and provide 2D and 3D ground truth joint positions, as well as foreground segmentation masks. Movements are chosen to be representative of infants in the first six months of life, and we divide the sequences into different levels of diffi-

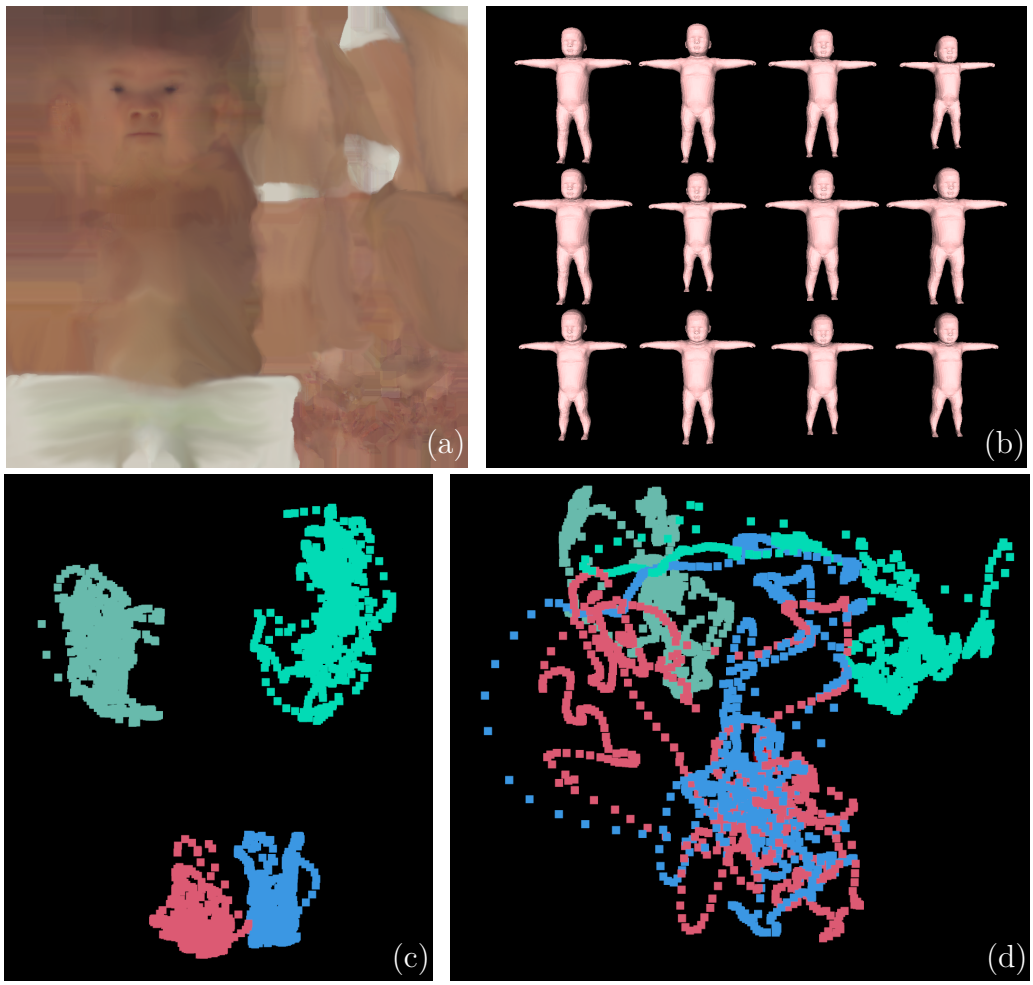


Figure 6.3.: (a) Sample of generated texture. (b) Generated shapes in T-pose. (c) Plotted joint positions from an "easy" sequence. Hand positions shown in light and dark green. Foot positions in red and blue. (d) Hand and foot positions for a "difficult" sequence. Color coding as in (c).

culty (see Fig. 6.3 (c) and (d) for examples): (i) easy: lying on back, moving arms and legs, mostly besides body, without crossing (sequences 1-4), (ii) medium: slight turning, limbs interact and are moved in front of the body, legs cross (sequences 5-9), and (iii) difficult: turning to sides, grabbing legs, touching face, directing all limbs towards camera simultaneously (sequences 10-12).

Different approaches utilize different body representations with different underlying skeletons. To properly compare these approaches, we add one frame in T-pose (extended arms and legs, cf. Fig. 6.3 (b)) for each sequence to calculate initial offsets between estimation and ground truth that can be used to correct for skeleton offsets.

The limitations of the underlying SMIL model include finger motions, facial expressions and hair. These are not represented by the model, which is why the hand is

fixed as a fist, and the face has a neutral expression.

6.4. Baseline Evaluation

We provide baseline evaluation for 2D pose estimation from the RGB part of our MINI-RGBD data set. The evaluation of 3D pose estimation from the RGB-D data can be found in Sec. 4.2.3 and Sec. 4.3.7.

2D Pose Estimation in RGB Images. We use a state-of-the-art adult RGB pose estimation system from OpenPose library[1] [Cao+17; Ope] as baseline for evaluation of the RGB part of the data set. To account for differences in skeletons between OpenPose and SMIL, we calculate joint offsets for neck, shoulders, hips, and knees from the T-pose frame (Sec. 6.3), and add these offsets to the estimated joint positions in every frame.

Error metrics. We apply the PCKh error metric from [And+14], which is commonly used for the evaluation of pose estimation approaches [Cao+17; Wei+16; Haq+16; Sci+17]. It denotes the percentage of correct keypoints with the threshold for correctness being defined as 50% of the head segment length. The SMIL model has a very short head segment (head joint to neck joint, cf. Fig. 6.2, (c) and (f)), which is why we present results using the full head segment length (PCKh 1.0), as well as two times the head segment length (PCKh 2.0) as thresholds. The head segment length for each sequence is calculated from the ground truth joint positions in the T-pose frame. Average 2D head segment length over all sequences is 11.6 pixels. We calculate the PCKh values for each joint for each sequence, and average numbers over all sequences, respectively over all joints.

OpenPose estimates 15 joint positions (nose, neck, shoulders, elbow, hands, hips, knees, feet) that we map to corresponding SMIL joints. Unlike SMIL, OpenPose estimates the nose position instead of head position. We add the model vertex of the tip of the nose as additional joint instead of using SMIL head joint.

Results. We display average PCKh per joint in Fig. 6.4 (a), and average PCKh per sequence in Fig. 6.4 (b). The mean average precision, i.e., the average PCKh over all joints and sequences, for PCKh 1.0 is 88.1% and 94.5% for PCKh 2.0. PCKh rates are very consistent over most body parts, with a slight decrease of PCKh 1.0 for lower limb joints, especially knees. Results for some body joints (e.g., nose, neck, shoulders, Fig. 6.4 (a)) as well as for some sequences (1, 2, 3, 6, 7, 11, Fig. 6.4 (b)) are close to perfect (according to the error metric). We observe largest errors when the limbs are directed towards the camera.

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose

OpenPose has reportedly obtained impressive results on pose estimation of adults [Cao+17], and confirms these on our synthetic infant data set. Being trained on real images of unconstrained scenes, the results further validate the high level of realism of our data set, but also show how challenging the data is, and that there is still room for improvement, e.g., sequences 9, 10, 12.

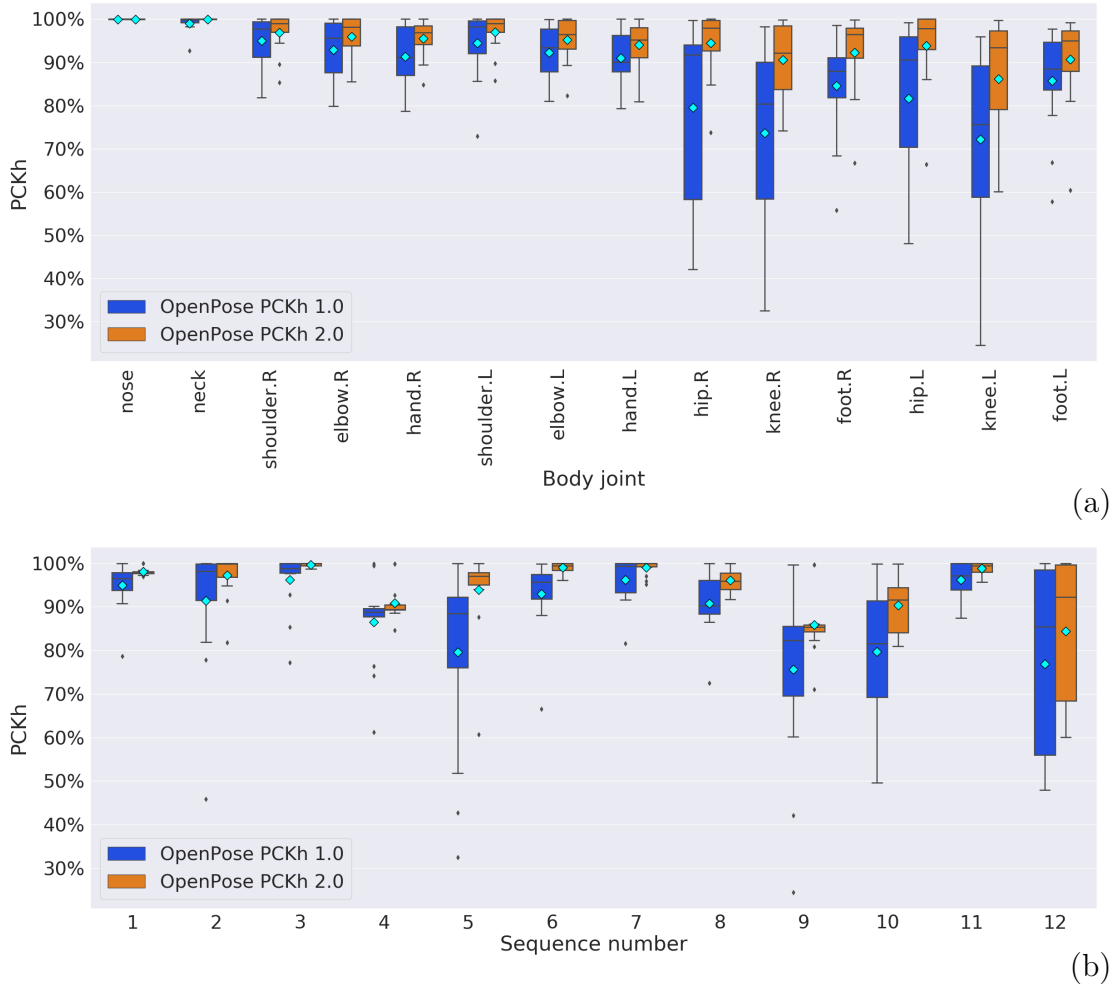


Figure 6.4.: RGB evaluation. Results for 2D pose estimation from OpenPose library. (a) Percentage of correct keypoints in relation to head segment length (PCKh) per joint. For *PCKh 1.0* a keypoint is considered correct if the distance to ground truth is smaller than one time the head segment length, and for *PCKh 2.0* if the distance is smaller than twice the head segment length. (b) PCKh per sequence. Cyan colored diamonds depict the mean value.

7. Towards Automated Medical Motion Analysis

In computer vision literature, the term motion analysis often is used as a synonym for recognizing an activity class from videos [Moe+06; Ye+13]. In our context, motion analysis refers to assigning a label or number that is representative of a medical assessment to a movement sequence, e.g., CP vs. no CP, or a GMA rating from one to ten.

Generally, an intermediate step between capturing motions and training a classifier/regressor is the calculation of motion features on which the training is based.

Extracting features from motion sequences for motion analysis can be approached in different ways.

- *Human interpretable measurements*: These include measurements like joint angles, velocities, or range of motion. These can be easily understood by humans, and statistics over these measurements can describe and summarize motion sequences.
- *Features that aim to resemble relevant motion properties*: These features rely on observations of specific phenomena and the assumption that these are medically relevant. Features previously used are, e.g., a stereotypy score [Kar+12], jerk index [Kan+14], or frequency analysis [Rah+15a].
- *Learned features*: Machine learning techniques can be used for embedding a sequence of captured motions in a latent space to separate different kinds of motions. These features are hard to interpret since they are automatically learned. The learning is guided by an objective function, which e.g. tries to group similar motion patterns close to each other in the latent space and to separate dissimilar patterns.

In the next section, we show how human interpretable motion parameters can be used to distinguish between different diseases. These parameters may provide quantitative information for doctors to back up a diagnosis or to measure progress over time. Researchers have tried using such parameters for CP detection, with varying success. One of the problems in CP detection is that it can have very diverse symptoms and different levels of severity. A parameter that seems valid for detecting CP in one infant may not work for another.

7.1. Human-interpretable Motion Parameters

Parts of this section were published as [Hes+17b].

We calculate motion parameters in this section from the body poses captured with the method we described in Sec. 4.3. The benefit of motion parameters introduced in this section is two-fold. First, the parameters provide objective measurements and can serve as an assistive tool for neurological examination. Second, they describe movement characteristics with the capacity of highlighting possible impairments and can therefore automatically detect infants in need of further examination.

7.1.1. Measurement-based Motion Parameters

The parameters are chosen to describe range, variability and symmetry of motion of arbitrary body parts. We conduct a preliminary study to show that captured measurements can provide helpful information to distinguish between different disease patterns. We measure values that depend on single body joints:

- velocity
- acceleration
- distance traveled
- distance to the table
- volume covered: volume of convex hull of joint positions of all frames
- percentage of frames in which motion is present ($velocity > threshold$)

Parameters related to head rotation are:

- number of head rotations
- percentage of frames in which the head is rotated towards the left/right side
- mean and standard deviation of head rotation angles

Measurements depending on multiple joints are:

- angles
- ratio between parameter values for left/right body side
- percentage of frames in which hands/feet/left side/right side/all extremities move together

We calculate motion parameters based on the above measurements. The parameter for traveled distance of an extremity d_tr_{extr} is the sum of Euclidean distances between joint positions in consecutive frames per minute:

$$d\_tr_{extr} = \frac{1}{t} \sum_{i=2}^{n} ||x_i - x_{i-1}||, \tag{7.1}$$

where *extr* denotes the extremity, t is the duration of the recording, x_i the 3D joint position in frame i, and n the number of frames in the recording. $||\cdot||$ denotes the Euclidean norm. The parameter describing the change in distance to the table of an extremity d_table_{extr} is defined by

$$d\_table_{extr} = \frac{1}{t}\sum_{i=2}^{n} ||dist(x_i, tp) - dist(x_{i-1}, tp)||, \tag{7.2}$$

with $dist(x_i, tp)$ specifying the distance of the joint position of the extremity to the table plane in frame i. The ratio $ratio\_lr(par)$ of a parameter *par* denotes the proportion of parameter values for joints of the left body side, compared to parameter values of the right side. This describes the factor by which the parameter value for one body side is larger than the value for the other side. A value of one implies symmetry of parameters for both body sides.

$$ratio\_lr(par) = \frac{max(par(l), par(r))}{min(par(l), par(r))}, \tag{7.3}$$

where $par(s)$ specifies the summed values of parameter *par* for left l and right r body side.

7.1.2. Clinical Study for Evaluation of Motion Parameters

We present 6 sequences of 5 infants, lying in supine position without external stimulation. All of them were awake and in a cooperative mood at the time of recording. Ethical approval (Ludwig-Maximilians-Universität Munich) and parents' written consent was obtained prior to recording. A subset of 9 parameters is selected from the full parameter set (Sec. 7.1), and evaluated on the recorded sequences (Fig. 7.1, Fig. 7.2, and Fig. 7.3).

Definitely healthy patient (DH_1) is a former preterm infant, born at 35+4 weeks of gestational age (WGA). At the time of recording, the patient is 14 weeks of corrected age. The patient shows head and trunk position close to midline, visually exploring the environment with normal head movements to both sides. All limbs move independently from each other and there are no stereotypical movement patterns. Fidgety movements are observed, representing age-corrected adequate gross motor development (RGB video - data not presented). Fig. 7.1 (DH_1) shows comparable spontaneous activity of left and right upper and lower extremities with ratios close to 1. Head rotation intensity and frequency is unremarkable (Fig. 7.2). The amount of movement is higher in lower than in upper limb (Fig. 7.3).

High Risk patient 1a (HR_1a) is a former preterm infant, born at 25+2 WGA. At the time of recording the patient is 12 weeks of corrected age. Head and trunk are in midline position with a slight rotational predominance to the right. No significant symptoms of abnormal gross motor development besides a mild positional

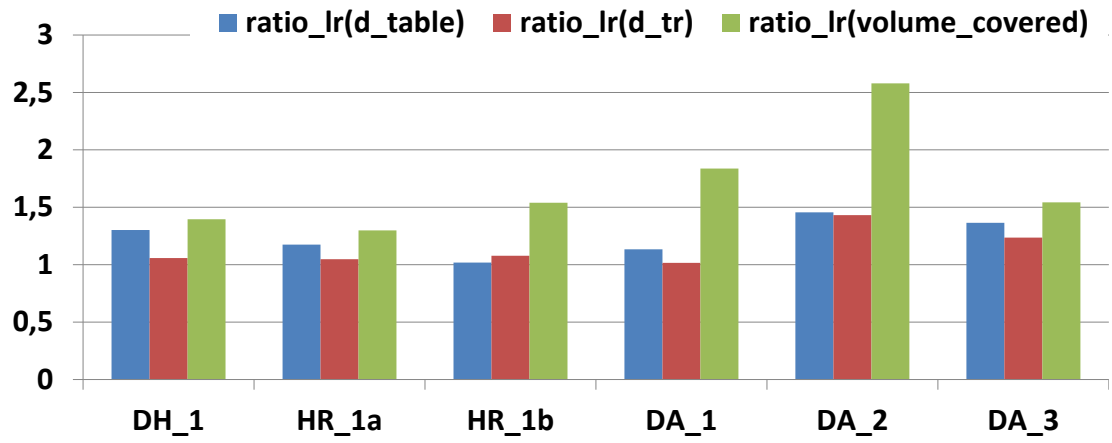


Figure 7.1.: Evaluation of motion parameters. Ratios between body joints for left and right body side for parameters distance to table (*ratio_lr(d_table)*), distance traveled (*ratio_lr(d_tr)*), and volume covered (*ratio_lr(volume_covered)*). DH_1 denotes a sequence of a definitely healthy infant, recordings HR_1a and HR_1b stem from the same infant from a high-risk population at different ages. DA_1-3 denotes recordings of infants with definitely abnormal motor development.

asymmetry of head rotation towards the right side are observed in RGB video (not presented). Fig. 7.1 (HR_1a) displays no significant lateralization of spontaneous arm or leg movements. Head rotational parameters are without remarkable differences (Fig. 7.2), and a high amount of motion in all limbs represents a very active baby (Fig. 7.3).

High Risk patient 1b (HR_1b) is the same patient as HR_1a, but at follow-up 4 weeks later. The previous preference of head rotation to the right has decreased to almost midline (RGB video - not presented). Fig. 7.1 (HR_1b) displays a mild lateralization of volume covered by spontaneous movements comparing left and right sides. Head movements have increased (Fig. 7.2), while limb movements have decreased compared to the prior analysis, with again predominance of the lower limbs, representing a less active infant (Fig. 7.3).

Patient 1 with definitely abnormal motor development (DA_1) is a term-born infant, who is 14 weeks of age at the time of recording and has an obvious cognitive impairment. A neurological examination resulted in diagnosis of a genetic syndrome including muscle weakness and muscular hypotonia. The infant shows an apparent lack of head rotation which is caused by the weakness of neck muscles and lack of interest in the environment (Fig. 7.2, DA_1). The head cannot be held in the truncal plane, leading to spontaneous positioning to the left (RGB video - not

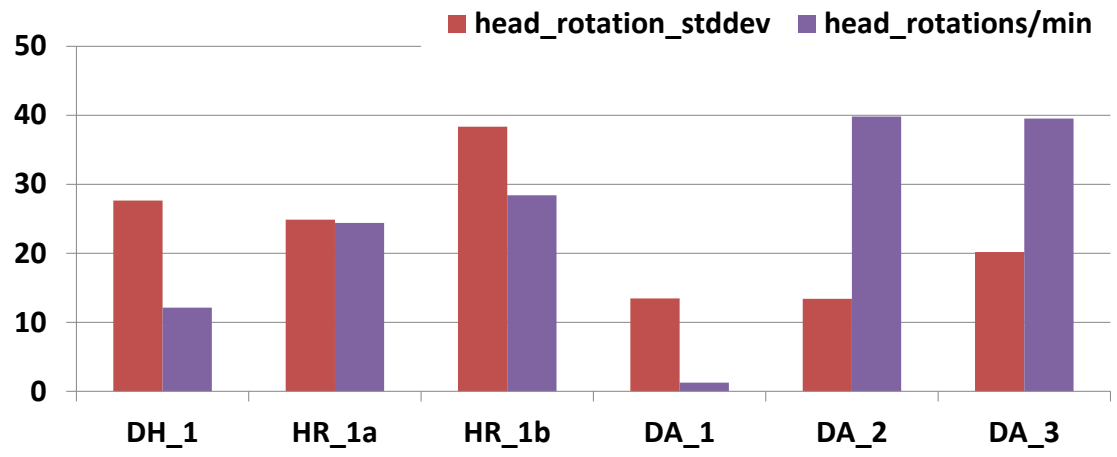


Figure 7.2.: Evaluation of motion parameters. Head rotation parameters per recording. Standard deviation of angles in degrees, number of head rotations per minute. DH_1 denotes a sequence of a definitely healthy infant, recordings HR_1a and HR_1b stem from the same infant from a high-risk population at different ages. DA_1-3 denotes recordings of infants with definitely abnormal motor development.

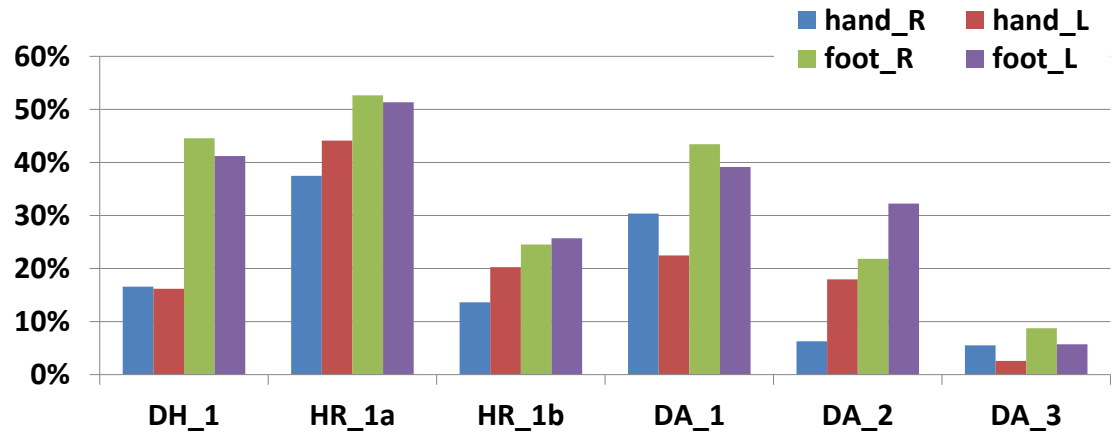


Figure 7.3.: Evaluation of motion parameters. Percentage of frames in which motion is present, indicated for each of the extremities. DH_1 denotes a sequence of a definitely healthy infant, recordings HR_1a and HR_1b stem from the same infant from a high-risk population at different ages. DA_1-3 denotes recordings of infants with definitely abnormal motor development.

presented). The asymmetrical tonic neck reflex limits the mobility, especially of the right arm, and causes a large difference in volume covered by left and right body side (Fig. 7.1). Weakness is predominantly present in the head, neck, and arms, while the legs are able to lift off the ground much easier, with again asymmetric distribution (Fig. 7.3).

Patient 2 with definitely abnormal motor development (DA_2) is a term born infant with generalized hypotonia due to a genetic syndrome, who is 34 weeks of age at the time of recording. The most obvious impairment is muscle hypotonia, but the infant has a good (age-adequate) mental state. Similar to DA_1, there is a significant lateralization in spontaneous movements of the upper limbs (Fig. 7.1). In contrast to DA_1, head movement is much better, as neck strength is sufficient to rotate the head, and the patient is interested in visually exploring the environment (Fig. 7.2). A predominance of motions of the lower limbs and the left body side is indicated in Fig. 7.3, but with a remarkably reduced amount of motion compared to DA_1.

Patient 3 with definitely abnormal motor development (DA_3) is a former preterm infant, born at 23+4 WGA. At the time of recording, the patient is 13 weeks of corrected age. Cranial MRI demonstrated intracranial hemorrhage III° on both ventricles, as well as signs of bilateral periventricular white matter lesions, leading to the diagnosis of bilateral spastic cerebral palsy. The muscle tone is generally increased on both upper and lower limbs. Head control is reduced, but possible, while the trunk is hypotonic. The patient is not able to open his hands actively and shows significantly reduced spontaneous, and stereotypical, generalized movement patterns of the upper and lower limbs (RGB video - not presented). Although there is hardly any spontaneous activity of the patient (Fig. 7.3), there are no significant differences in movement distance and volume covered when comparing left and right upper limb (Fig. 7.1). Head rotations are within normal ranges (Fig. 7.2).

7.1.3. Discussion

The motion parameters in this section may provide helpful information to backup doctors' observations with measurements. The number of subjects is too small to derive general assumptions, or to consider training a classifier on such diverse data. We showed that measured motions can visualize differences between diseases, but also provide quantitative measures, that could be used to track the development of patients over time.

As of today, clinical assessments have been designed with human observational skills in mind. Now that a method is available to reliably measure motion over time, assessments relying on such measurements could be developed.

7.2. Towards Automated General Movement Assessment

Medical knowledge allows doctors to define parameters that are relevant for diagnosing or classifying a disease. Classical medical assessments judge the execution of different predefined tasks. These tasks are defined in a way that an observer can either classify the performance in a repeatable manner, e.g., the *Movement Assessment Battery for Children – second edition* [Hen+07], or that the time duration to complete the given task can be used as performance measure, e.g., the *Timed Up and Go* test [PR91]. A good assessment provides general applicability, repeatable results, and strives for invariance to human perception. The advantage of such assessments is the option of comparing the results to a healthy population as a reference.

We have described the general movement assessment in detail in Sec. 2.2, but briefly recap the most important properties that will be relevant in this section. GMA is performed up to the age of four months, before infants start using directed movements. Infants can not follow instructions, which eliminates the option to use predefined tasks. Instead, GMA assumes that, when the infant is in an active, wakeful state, the brain continuously generates seemingly random movements. The generated patterns show less *variation*, i.e., the infant has a smaller **repertoire** of movements, and **complexity** if there is an impairment of the brain [HA04]. GMA assesses the quality of spontaneous infant movements with respect to these two properties that aim to describe the relevant features for visual Gestalt perception. The movement quality of a recording is assigned to one of the classes[1] *definitely abnormal* (DA), *mildly abnormal* (MA), *normal suboptimal* (NS), or *normal optimal* (NS). The class DA is associated with a high risk of CP [HA04].

In this section, we present a method using a motion sequence to predict if an infant shows definitely abnormal motion quality and therefore has a high risk of CP. To achieve this, we divide motion sequences into small snippets of *motion words*, which we embed in a learned feature space. The feature space groups motion words based on similarity of motion and GMA rating. This allows us to evaluate the quality of motion words from a new sequence based on their location in feature space. The statistics of all motion words of a sequence enable us to calculate *complexity* and *repertoire* features to classify the sequence as either *DA* or *not-DA*.

7.2.1. Related Work – Automated General Movement Assessment

Researchers have translated these subjective descriptions of movement quality into mathematical formulations [Mei+06; Kar11]. They define motion features similar to

[1] Over the course of this thesis, we have used the terms "GMA rating" and "GMA class" relatively interchangeably. In this section, we will use the term "GMA rating" to refer to a 1-10 scale, and the term "GMA class" to refer to one of the classes DA, MA, NO, NS.

the description of GMA in literature, e.g., spatial distribution of limb movements, activity of joints, and limb velocities and accelerations. However, experimental results for this type of measurement could not be reproduced with different data sets [Kar11].

[Kar+12; Kan+14; Rah+16] introduce higher-level features, like stereotypy, jerkiness of motions, and movement frequency analysis. As discussed in Sec. 2.4, promising results have been achieved but none of the approaches has prevailed as *the* valid and reliable method for CP detection. What is common in all presented studies is the small number of infants who are affected by CP. The maximum number in a single study is 16, with the maximum number of healthy subjects being 127 [Orl+18]. A classifier trained on such a small number of samples only has limited general validity. More literature on automated CP detection is discussed in Sec. 2.3 in detail.

Most studies analyze motion sequences that are captured before the infants' age of *four months*, but aim to predict the CP outcome at the age of *two years*, when a reliable diagnosis can be established. GMA has proven to be the most reliable tool for detection of CP at a young age. The developing brain undergoes large changes in the first two years of life, which includes the possibility of "outgrowing" problems that were caused at an early age [NE82]. The diagnosis at two years may also be influenced by pathologies that were developed after GM age. For this reason, we do not predict the CP outcome (at two years age) from sequences that are recorded at two to four months age. We want to assess the state of the infant *at the time of recording*, and therefore predict the GMA classification.

A very recent preprint[2] targets to identify infants with abnormal GM quality from four body-worn accelerometer signals [Gao+19]. Similar to our approach, they split motion sequences into smaller snippets and use a "weak labeling" approach by assigning each snippet the label of the full sequence, which can be *abnormal* (AN), *neutral* or *typically developing* (TD). For the creation of snippets, they use a sliding window with a duration of one second without overlap. A nearest neighbor method determines the GMA rating label for each snippet of a new sequence, and the labels are accumulated. Neutral movements are considered to be irrelevant and therefore ignored. The numbers of abnormal and typically developing snippets are aggregated in a score, which determines the class label of a sequence based on an experimentally specified threshold.

The data set includes 21 typically developing infants and 13 infants with perinatal stroke, showing abnormal motor development, who have been recorded multiple times over the range of six months. The initial set contains 95 TD and 60 AN sequences, of which some are removed due to recording issues. Some very long sequences are split into multiple shorter ones, and the final data set consists of 161 samples. A sensitivity of 70% and a specificity of 87% is reported using (random) 10-fold cross validation.

[2]This paper was uploaded shortly before submission of this thesis, which is why it is not analyzed in more detail in Sec. 2.3.1.

Two issues remain: i) Due to random splitting of the data set in the cross validation evaluation it is very probable that samples from the same sequence (long sequences that have been split) appear in both training and test set. ii) It is unclear if infants have subject-specific movement repertoires during the first six months of life. Due to the comparatively high number of *recordings* in relation to *infants* in the data set, it is inevitable that each infant in a test sequence is also included in the training set, however at a different age. *If* subject-specific motions are consistent over time, this may bias results.

7.2.2. Choice of General Movement Assessment Variant

There are two variants of GMA with slightly different scoring properties (cf. Sec. 2.2.1). *Prechtl's GMA* sets focus on presence, absence or abnormality of GMs, especially in "fidgety" age [Pre+97]. In addition, the occurence of cramped-synchronized GMs is a marker of abnormal movement quality.

Hadders-Algra proposes a more general and continuous scale of movement quality [HA04]. Depending on the complexity, variation and, to a lesser degree, fluency of movements (cf. Tab. 2.3), a recording is assigned to one of four motion quality classes, which can be augmented with a one to ten scale for finer granularity. The lowest class, *definitely abnormal* (DA), hints at a high risk of CP. For infants assigned to one of the other classes no clear prediction can be made, although mildly abnormal (MA) movement quality is associated with a higher risk for minor neurological dysfunction [HA+04]. For this reason, a further division into high/low CP risk can be made by evaluating DA (1-3) versus the remaining classes MA, NS, and NO (4-10), which we denote as *DA vs. not-DA*. We additionally examine how well a classifier can divide abnormal (DA, MA) from normal (NS, NO) motion quality, which is indicated by *DA+MA vs. NS+NO*.

We apply GMA according to Hadders-Algra, since the more continuous nature of GMA ratings helps to achieve a more fine-grained differentiation of movement quality. In order to classify certain types of GMs similar to Prechtl's GMA, we would need to identify beginning and end of each GM in all sequences. Unfortunately, the annotation of a large number of sequence requires huge effort by GMA experts and, to our knowledge, no such data set is available.

7.2.3. Classifying Sequences using a Learned "Motion Word" Feature Space

Deep learning approaches have shown tremendous success in predicting labels for given input data, e.g., identifying objects in images [Rus+15]. However, the use of such methods for automated CP detection/GMA has been reported to be unsuccessful [Kar11; Gao+19], which has been attributed to the diversity of motions inside a given sequence in combination with very few samples per GMA class.

Instead of analyzing all frames of a motion sequence at once, we intend to find a more fine-grained representation of a sequence that is characteristic of GMA.

Bag of "Motion Words". The task of activity/action recognition shows several similarities to our problem – it aims to assign a class label to a motion sequence, e.g., "walking" or "jumping". Recently, Aristidou et al. introduced a method for activity classification [Ari+18b], which is based on the *Bag-of-Words* (BoW) idea [Wit+99]. Initially developed for compression and indexing of text documents as well as images, the concept can be transferred to motions sequences. In our case, a word is represented by a set of body joint angle values. We divide a motion sequence into **motion words**, which are small, continuous snippets of predefined duration. The size of a motion word is defined as $\#joints \times duration$.

The BoW approach creates a *dictionary* from a corpus of documents. A *document* is represented by the histogram of words it contains. Frequently occurring words are less discriminative than rarely occurring words, which is why the histogram is re-scaled with *term frequency - inverse document frequency* (tf-idf) weighting, which is proportional to the word frequency in the motion sequence and inversely proportional to the frequency in the corpus of training words.

The term frequency is given as

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \tag{7.4}$$

with t denoting the *term*, i.e., the current word, d the current *document*, and $f_{t,d}$ the *frequency* of t in d.

The inverse document frequency is defined as

$$idf(t,D) = \log \frac{N}{1 + |\{d \in D : t \in d\}|}, \tag{7.5}$$

where D denotes the set of all documents, $N = |D|$ the number of all documents, and $|\{d \in D : t \in d\}|$ the number of documents in which t occurs. To avoid division by zero, 1 is added to the denominator.

The tf-idf weighting is calculated as

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D). \tag{7.6}$$

Aristidou et al. apply the method to the task of activity recognition with the underlying assumption that similar activities contain similar motions that appear frequently [Ari+18b]. Each motion sequence is represented by a *motion signature* that allows to distinguish different activity classes. Motion words of a sequence are *embedded* in a learned feature space with a triplet loss CNN. Each triplet contains an anchor, a positive sample that is similar to the anchor, and a negative sample that is dissimilar

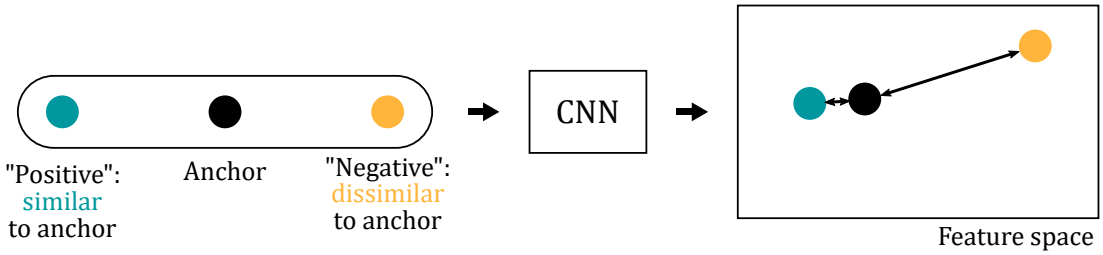


Figure 7.4.: A triplet of *motion words* is formed by selecting "positive" and "negative" samples for an anchor word. Positives are similar to the anchor, negatives are dissimilar. A CNN is trained with a triplet loss to create a feature space in which positives are close to the anchor and negatives are far away. This way, motion words with similar characteristics can be grouped.

to the anchor, as visualized in Fig. 7.4. The CNN learns to place the anchor words close to positives and far from negatives, i.e., random samples from other sequences. The similarity between motion words is calculated with the *dynamic time warping* (DTW) measure [MÖ07]. The CNN takes a motion word as input and generates a vector of predefined size, which is usually smaller than the size of the motion word.

K-means clustering of motion word embeddings produces a dictionary of size k, and each cluster is represented by a *motion motif*, which is defined as the cluster center. These motifs represent the most common and discriminative words. The motion signature of a new sequence is generated by assigning its motion words to motion motifs and creating a tf-idf-weighted histogram of word occurrences. A K-Nearest Neighbor classifier using the Earth Mover's Distance as a motion signature similarity measure predicts the class label of a new sequence [Ari+18b].

Learning features that represent GMA properties. Activities are defined by motions, and the motions are similar for different persons performing the same activity. The task of estimating the GMA class from motion sequences is different. Spontaneous movements of infant are not directed, and follow no recognizable pattern. GMA is not defined by certain movements, but rather by complexity and variation [HA04]. **Complexity** refers to the spatial variation over a sequence – high complexity is described as "frequent changes in direction of the participating body parts". *Variation* denotes temporal variation – across time, the infant produces continuously new movement patterns. A limited **repertoire** of movements hints at an impaired brain.

Therefore, we use the BoW approach not to create motion signatures per sequence, but rather to build a dictionary of motion words along with their characteristics. In an *ideal* feature space for GMA, two properties for motion words are satisfied:

i) *Words that are close to each other are similar.* This implies that the occurrence of two close words does not "increase" the observed motion repertoire, whereas the

occurrence of two distant words does. This property requires a good measure for similarity. Dynamic time warping has been shown to be successful [MÖ07], but it has been reported that a learned feature space can separate different types of motion sequences better than original DTW while reducing the time to compute similarities by up to 80% [Ari+18b].

ii) *Nearby words share the same level of complexity.* This means that the neighborhood of a word can be used to determine its level of motion complexity. To satisfy this property, a large set of annotations of motion word complexity are required. No such data set exists, and the annotation involves immense effort by trained experts. To overcome the lack of data, we make use of what Gao et al. call "weak labeling" [Gao+19]. We label all motion words with the GMA rating of the sequence they belong to. We make the assumption that sequences contain motion words that are characteristic of the associated rating, but at the same time contain motion words that will also occur in sequences with different ratings. The number of good motion words is expected to be higher in highly rated sequences than in sequences with low ratings, and vice versa. Therefore, we propose to find words that are most similar to the input word, together with their label. We expect the accumulated labels to reveal information about the GMA class of the input sequence.

Training a CNN with triplet loss. Similar to [Ari+18b], we train a CNN with triplet loss to learn a feature space with the goal of capturing the complexity and repertoire properties we defined above. Positive words in a triplet are chosen from the same sequence as the anchor and are either temporally close (without overlap) or have a small DTW distance, which is why we denote this type of triplet as $\mathbf{trip_{seq}}$. The restriction to the same sequence is applied, because the calculation of pairwise DTW similarities for all words of all sequences is computationally intractable. A sample for two motion words with a small DTW distance can be found in Appendix C. Negative words are chosen randomly from the rest of the data set. For each anchor, nine triplets are formed, with four temporally close positives and five positives based on the corresponding DTW distance.

An inspection of the learned feature space reveals that the CNN learns to closely cluster motion words of the same sequence. However, it fails to cluster motion words from different sequences that have the same rating, solely based on motion similarity. We visualize the distribution of a subset of all motion word embeddings in the learned feature space in Fig. 7.5 by projecting it to a two-dimensional space using t-SNE [MH08]. The color of the motion word embeddings denotes the corresponding sequence ID. Comparing the training samples (top left) and test samples (top right) shows that many small groups of the same color are formed[3], with less pronounced clustering in the test samples. The middle column shows the union of training and test samples. We color the embedded motion words according to their *ratings* in

[3]The number of sequences is higher than the number of easily distinguishable colors, which results in few sequences sharing the same color.

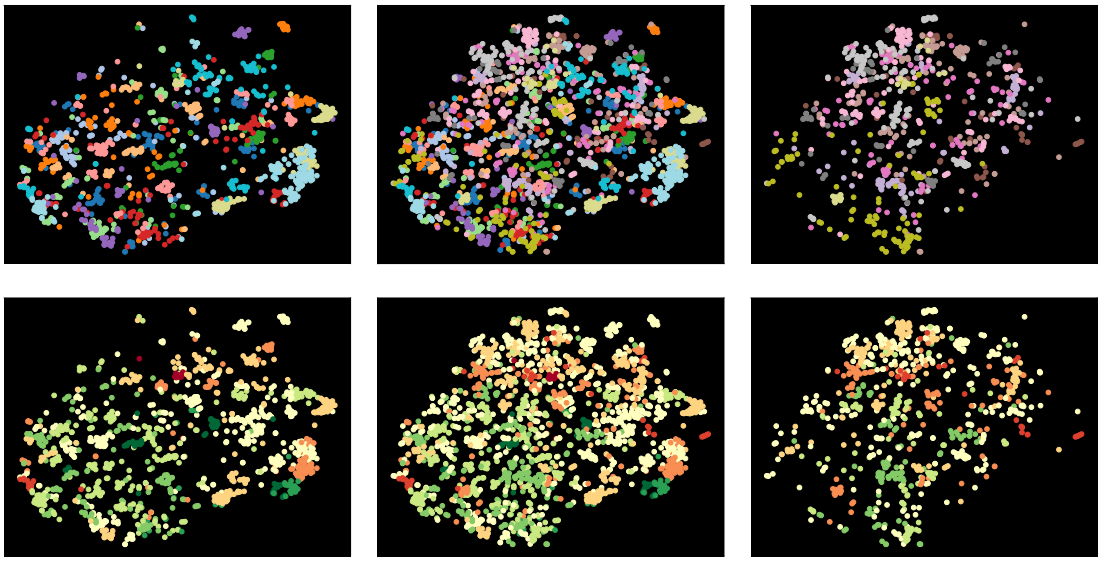


Figure 7.5.: Visualization of a subset of motion words embedded in the feature space that was learned using triplet loss $\mathbf{trip_{seq}}$ similar to [Ari+18b] with positives exclusively from the same sequence. The embeddings are projected to two dimensions using t-SNE [MH08]. Motion words with the same color share the same label. **Left**: Training samples only. **Middle**: Training and test samples. **Right**: Test samples only. **Top**: Color relates to sequence ID (multiple sequences may share the same color). One can observe many small clusters of words of the same *sequence* and therefore of the same GMA rating. **Bottom**: Color relates to GMA rating. From dark red (lowest rating) to dark green (highest rating). There is no obvious structuring of *ratings* inside the feature space.

the bottom row of Fig. 7.5. In the training samples (bottom left), small clusters of similar ratings exist, but when comparing them to the sequence based coloring it becomes obvious that the grouping of ratings seems to be due to the fact that all words of a sequence share the same label. The concept of clustering based on motion similarity does not suffice to induce an ordering of ratings in feature space. Since we are dealing with unconstrained motions, the probability of two motion words being very similar and therefore ending up in the same location in the feature space is small. The property of sequence "affiliation" dominates over inter-sequence word similarity.

In order to promote a closer placement of motion words with *similar ratings from different sequences*, we aim to relax this property. We create a new type of triplets by adding randomly selected words from sequences with *similar* GMA ratings as *positives*, and randomly selected words from sequences with *different* ratings as *negatives*. We call this type of triplet $\mathbf{trip_{GMA}}$. For each anchor word, we add five new triplets, but at the same time reduce the number of temporally close positives from the same sequence, to soften the sharp boundaries of sequence clusters. We

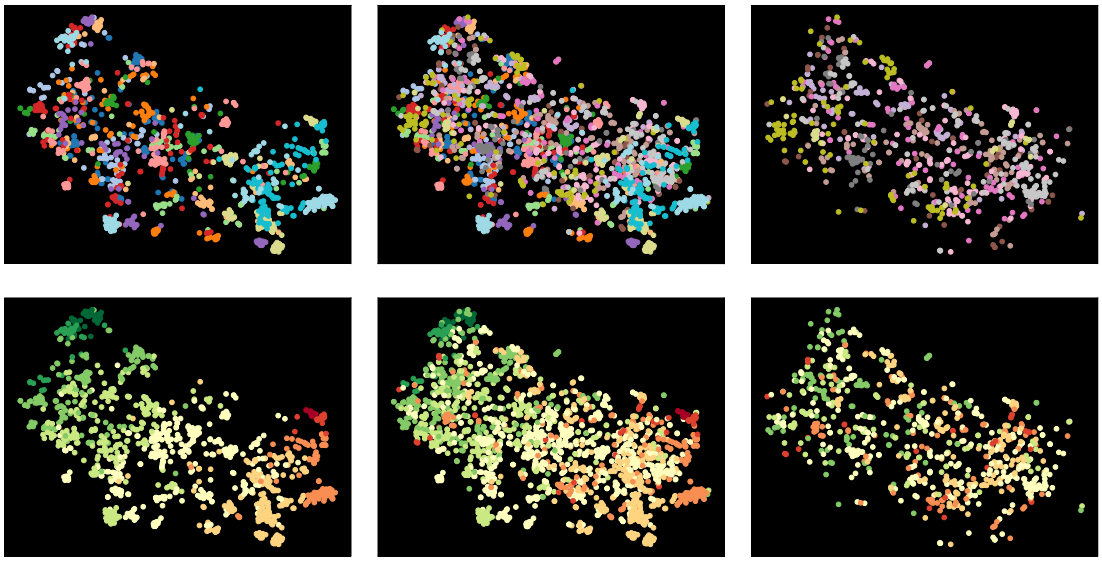


Figure 7.6.: Visualization of a subset of motion words embedded in the feature space that was learned using our new triplet loss $\mathbf{trip}_{\mathbf{GMA}}$ with added positives from different sequences with *similar* ratings, and negatives from sequences with *different* ratings. The embeddings are projected to two dimensions using t-SNE [MH08]. Motion words with the same color share the same label. **Left**: Training samples only. **Middle**: Training and test samples. **Right**: Test samples only. **Top**: Color relates to sequence ID (multiple sequences may share the same color). The *sequence*-based clustering appears less organized in both training and test set.**Bottom**: Color relates to GMA rating. From dark red (lowest rating) to dark green (highest rating). The CNN has placed training samples according to their *rating*. The ordering in the test set is less pronounced.

visualize in Fig. 7.6 training and test sample embeddings, colored either by sequence ID or GMA rating.

In the *training* samples (top left), the *sequence*-based clusters are still visible. The *rating*-based coloring (lower left) shows a clear separation of different ratings. Green samples on the left depict high ratings, red samples on the right denote low ratings. The structure in the *test* samples is less obvious. This may be in line with our assumption that any sequences may contain movements of differing quality, and that the overall statistics matter, i.e., highly rated sequences contain *more "good" words*, and sequences with low ratings contain *more "bad" words*. An experimental evaluation is necessary to determine if the CNN has learned to group motion words that are similar and are annotated with the same rating label, or if the focus on separation of ratings is too strong and dominating the similarity-based grouping.

The triplets serve as input for a CNN that is based on the inception model [Sze+15] and the FaceNet architecture [Sch+15][4]. After training, we project the samples into

[4]We adapt an implementation from `https://github.com/krasserm/face-recognition`.

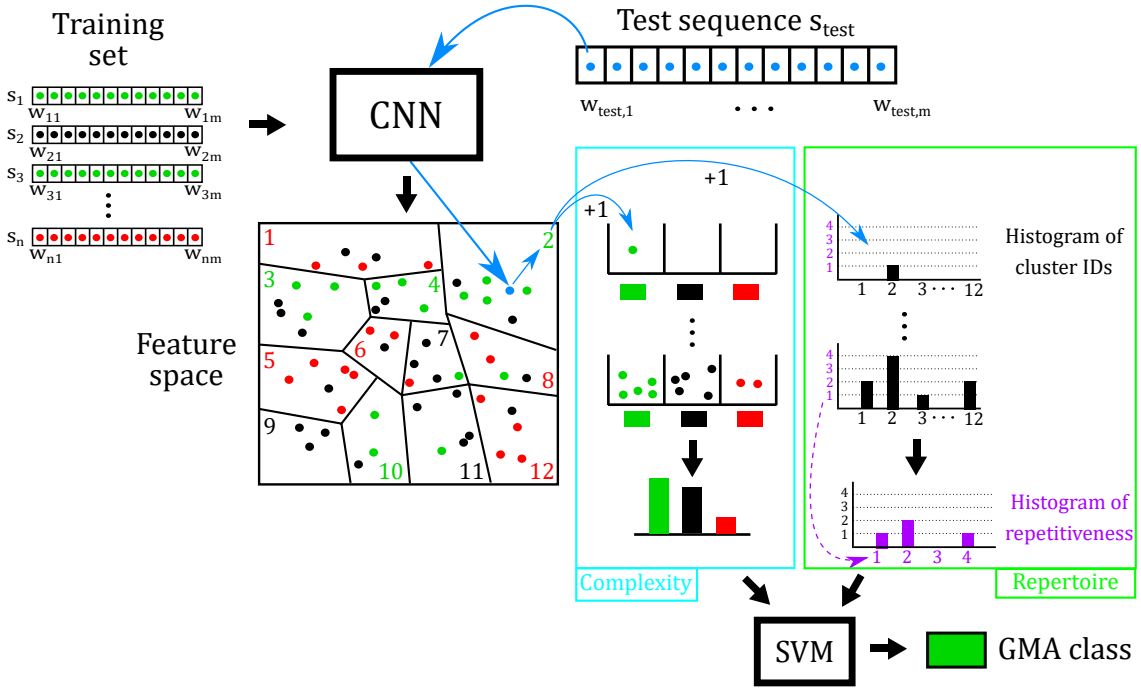


Figure 7.7.: Overview of the computation of complexity and repertoire features. A CNN learns a feature space of motion words from the training set. Training words are projected to the feature space with their GMA rating labels (shown as red, green and black). K-Means clustering divides the space into a discrete dictionary and associates a GMA label with each cluster (here k = 12). Words $w_{test,i}$ with unknown GMA label (blue) of a test sequence are projected to the feature space, and the dictionary cluster determines the label of the input word. All word labels of the test sequence are accumulated in a "complexity" histogram. Additionally, a "histogram of repetitiveness" is created from the shown repertoire of motion words in the sequence. These two normalized histograms serve as input to an SVM classifier that assigns a GMA class label to the input sequence.

the new feature space. We perform k-Means clustering on the embedded samples to transform our feature space into a discrete motion word dictionary. Additionally, each cluster is assigned with a GMA rating. This label is determined by finding the rating with the highest normalized frequency among the words inside the cluster.

Creating features to represent complexity and movement repertoire. By embedding a new motion word in the feature space, we can not only assign a word ID from the dictionary to the new word, but also a GMA rating. The word IDs over all words of a sequence provide a representation of the repertoire of infant movements, and the accumulated GMA labels serve as a measure for motion complexity.

We provide an overview of our method and the computation of complexity and repertoire features in Fig. 7.7. First, we project all training words to the learned

feature space together with their GMA rating labels. We apply k-Means clustering to divide the space into a discrete dictionary, and associate the dominant GMA label with each cluster. Words with unknown GMA label of a test sequence are projected to the feature space, and we determine the word ID and the GMA class label for the input word from the dictionary. All GMA word labels of the test sequence are accumulated in a *complexity histogram*. The *histogram of repetitiveness* is created from the shown repertoire of motion words in the sequence. It contains statistics of word repetitions. Additionally, the ratio of the relative repertoire size, i.e., the number of distinct words in relation to the number of words in the sequence, is included. These two normalized feature vectors serve as input to an SVM classifier for assigning a GMA class label to the input sequence.

[Ari+18b] use a motion sequence signature for classification, which is represented by a normalized histogram of dictionary word IDs. Our data set is small and, due to the unconstrained nature of infant motions, very diverse. We need a rather large vocabulary to pick up fine grained movements, which results in a large and sparsely populated motion signature. The motion signature approach seems to be more suited for tasks in which a few, discriminant motion words occur repeatedly for each class one wants to distinguish.

7.2.4. Experiments

Data. We recorded 124 sequences of 120 different infants with a Microsoft Kinect V1 (107 sequences) and a Kinect V2 (17 sequences). Four infants in the data set have been recorded twice at different ages. When splitting the data into training and test set, we ensure that sequences of the same infant are only included in either training or testing set. As described in Sec. 3.2.4, the Kinect V2 recordings are subject to strong noise, which distorts the body shape and is probably caused by the close proximity of foreground and background. For this reason, we excluded these sequences during learning the SMIL model (Sec. 5.3), since we wanted to use only realistic shapes for training. Even though the *shapes* of the registered model may look unrealistic, we are able to correctly capture the *motion* from the noisy sequences and include them in our experiments.

All sequences have been rated according to the GMA *variant of Hadders-Algra*. The rater who performed the GM assessment for our data set *is* Mijna Hadders-Algra, who is the most experienced rater for her variant of GMA worldwide. She has been teaching GMA in courses for decades, and has defined the gold-standard videos with ratings that are used for re-calibration and practice. For this reason, we consider the GMA ratings of our sequences to be highly reliable.

As noted previously, the overall number of infants affected by CP/showing abnormal motor development is small, which is why our data set is very unbalanced w.r.t. the number of sequences per class/rating as can be seen in Fig. 7.8. The vast majority

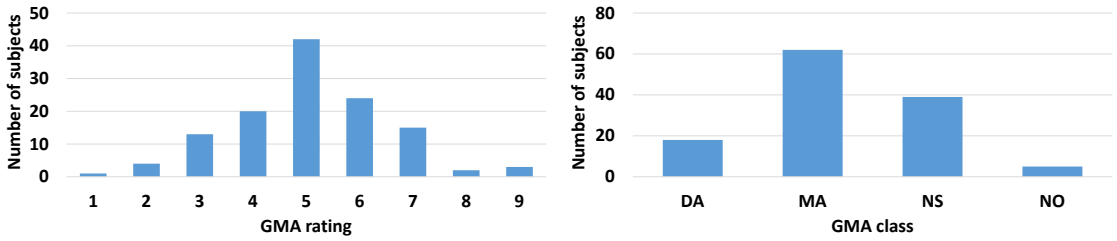


Figure 7.8.: Left: Distribution of GMA ratings/classes of 124 sequences in our data set. Left: GMA ratings from 1-10. Right: GMA classes. The vast majority has a rating in the range of mildly abnormal (scores 4-5) and normal suboptimal movement quality (6-7). The data set contains 18 infants in definitely abnormal class (1-3), who are considered to have a high risk of CP.

(101 recordings) are rated either as mildly abnormal (ratings 4-5) or normal suboptimal (6-7). The data set contains 18 infants labeled with the definitely abnormal class (1-3) who are considered to have a high risk of CP.

We capture motions from RGB-D sequences using the SMIL model, as described in Sec. 5.3. This gives us 72 angle values (3 global rotation angles + 23 joints * 3 degrees of freedom) for each frame of the sequence. As discussed in Sec. 5.6, SMIL does not capture fingers or toes, which is why we exclude the associated joints, together with one degree of freedom of one of the spine joints, which has been held fixed during the registration process. We further exclude the global rotation, since our focus is motion and not global positioning, which leaves us with 56 joint angles per frame. We divide the motion sequence into overlapping motion words of 32 frames length using a step size of 4 to reduce computational complexity. The overlap ensures that the DTW finds good matches, since we cannot assume that similar motions start at the beginning of a motion word. This results in motion words of size 56×32. Motion sequences may contain sections in which the infant does not move. These are irrelevant for GMA, since only the quality of motions is of interest, not the quantity. For this reason, we filter out words with an insignificant amount of motion. Our complete data set contains 112K motion words.

As evaluation metric, we provide sensitivity and specificity values for all experiments (cf. Sec. 2.1).

We design several baseline experiments to compare our approach, but also to analyze the proposed features for evaluation of infant motion. The baselines are related to either complexity or repertoire features, but do not include the training of a feature space. Similar motion words are identified by the Euclidean distance or dynamic time warping, and the entropy is used as an additional measure for motion complexity.

Baseline – complexity – k-nearest neighbor rating labels (B-kNN). We use a k-nearest neighbor approach to find for each input motion word the GMA rating

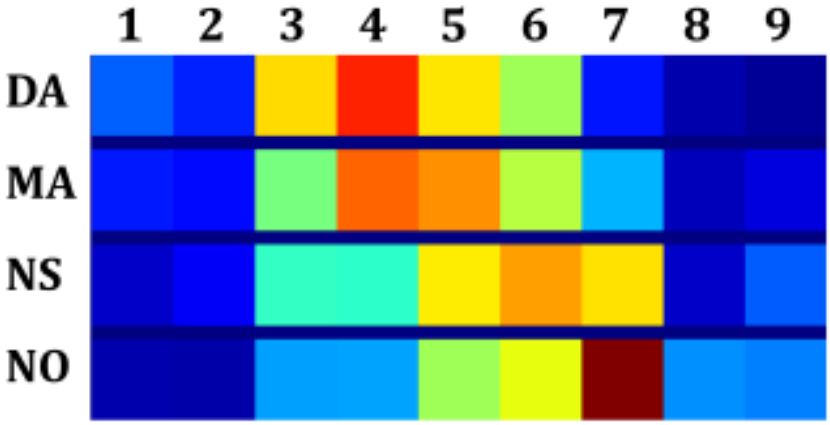


Figure 7.9.: Distribution of GMA ratings (x-axis) of motion word nearest neighbors with respect to GMA class of input sequence (y-axis), mean over all sequences per class. Colored fields illustrate the values of normalized histogram bins, ranging from high (dark red over orange to yellow) to low (dark blue). The maximum (red) seems to shift to the right with improving movement quality (from DA in top row to NO in bottom row).

label of the closest word in joint angle space (w.r.t. euclidean distance). The rating labels are accumulated in a histogram of size 10 and weighted with tf-idf. These feature vectors will serve as input for classification. The B-kNN complexity features are computed in the same manner as in our approach (cf. Fig. 7.7).

In order to explore the correlation between feature vectors of different GMA classes, we split the set of all feature vectors in groups, according to the GMA class label of each sample. We calculate the mean values inside each group w.r.t the histogram bins. The results in Fig. 7.9 support the assumption that abnormal motion sequences contain more movements from sequences with low ratings, and "normal" sequences a higher rate from sequences with better ratings. The classes definitely abnormal and mildly abnormal show relatively similar characteristics.

Baselines – complexity – entropy per joint (B-ENT). The entropy is a measure for uncertainty, and variants of entropy have been successfully applied to gait analysis as a measure of motion complexity [Cos+03]. One of our goals is to measure complexity, which is why we investigate the entropy of the motion sequences of our data set. We evaluate three different variants. We compute the entropy per joint over the full sequence (B-ENT$_{full}$) and use this feature vector as input to the SVM classifier. The second and third variant split the sequence into motion words of size 16 (B-ENT$_{16}$), respectively 32 (B-ENT$_{32}$), and calculate the entropy over all joints of each motion word. These values are accumulated in a histogram, i.e., the number of entropy values inside pre-specified ranges, and averaged over all joints.

Baselines – repertoire – dynamic time warping similarity (B-DTW). As noted above, it is not tractable to compute DTW distances between all motion words of the data set, and we only calculate the DTW similarity between words of the same sequence. We are lacking a similarity measure against other sequences and therefore can not collect rating labels from other sequences' words to accumulate in a histogram as in the complexity feature of our method. Instead, we use the computed intra-sequence DTW distances to calculate a measure for the motion repertoire. For each motion word of a sequence, we add the index of the closest word – according to DTW distance – to a list. We count the number of occurrences of each word index in the list and create a normalized histogram over these counts, which are input to an SVM. The B-DTW repertoire features are computed as in our approach (cf. Fig. 7.7).

Classification. In order to calculate meaningful evaluation results, the data set must be divided into training and test set. We visualize the training procedures with the data set splitting strategies in Fig. 7.10.

To handle the limited amount of training data, we perform *leave-one-out cross validation* (LOOCV) to repeatedly train the system on all but one samples, and test on the remaining sample. This implies a repetition of the training procedure n times with n being the number of samples in the data set. Training our CNN takes several hours, which is why the LOOCV scheme is too time-demanding. We split the data D_{ALL} into two subsets $Train_{CNN}$ and $Test_{CNN}$ and use $Train_{CNN}$ to train our feature space.

For the dictionary training with *k-Means clustering*, we apply LOOCV on $Test_{CNN}$, which is denoted by $D_{k-Means}$. We project motion words of all training sequences $Train_{k-Means}$, but also those of $Train_{CNN}$, to the feature space. This is necessary to provide enough training data to create a meaningful dictionary. We "look up" the words of the remaining sequence in $Test_{k-Means}$ in the dictionary and calculate repertoire and complexity features. Despite having a different dictionary for each test sequence, the computed features have the same reference, since our features rely on the motion word *statistics* and not on word IDs, which are related to a specific dictionary. Then, we switch the roles of $Train_{CNN}$ and $Test_{CNN}$ and repeat the procedure until we have generated features for all sequences of the data set D_{ALL}, to produce D_{SVM}.

We use the repertoire features, the complexity features, and a combination of both to train a *support vector machine* (SVM) classifier[5] on D_{SVM} with LOOCV. Before training, we perform a normalization of the data, which is also applied to the test data. We train the SVM to perform two different binary classifications, namely

[5]We use an SVM from the Python package sklearn `https://scikit-learn.org/stable/modules/svm.html`, with a radial basis kernel, and apply data balancing to account for different numbers of samples per class.

Table 7.1.: Results. We present sensitivity and specificity for baseline methods B-kNN (complexity) with k ∈{1, 5}, entropy histograms B-ENT (complexity), and B-DTW (repertoire). The number of frames defining the duration of motion words is denoted as *16* and *32* – *full* means that the entropy histogram is calculated over the complete sequence. $Ours_{\text{seq}}$ denotes our approach with triplet loss trip_{seq}, and $Ours_{\text{GMA}}$ with triplet loss trip_{GMA}. We specify the different types of features used for classification as "*(comp.)*" for the complexity features and "*(rep.)*" for the repertoire features. "*(both)*" represents a combination of both types. *Ours* is always trained with a word length of 32.

	DA vs. **not-DA**		**DA+MA** vs. **NS+NO**	
	Sensitivity	Specificity	Sensitivity	Specificity
B-1NN$_{16}$	50%	57%	**79%**	68%
B-1NN$_{32}$	39%	58%	78%	64%
B-5NN$_{16}$	39%	58%	74%	70%
B-5NN$_{32}$	44%	53%	75%	68%
B-ENT$_{16}$	**83%**	75%	64%	66%
B-ENT$_{32}$	67%	68%	53%	68%
B-ENT$_{full}$	56%	69%	74%	70%
B-DTW$_{16}$	**83%**	75%	60%	55%
B-DTW$_{32}$	78%	69%	76%	64%
Ours_{seq} (comp.)	33%	64%	68%	70%
Ours_{seq} (rep.)	67%	43%	44%	59%
Ours_{seq} (both)	39%	65%	65%	68%
Ours_{GMA} (comp.)	56%	68%	64%	**75%**
Ours_{GMA} (rep.)	39%	73%	35%	**75%**
Ours_{GMA} (both)	39%	**76%**	64%	**75%**

"DA versus not-DA", and *"DA+MA versus NS+NO"*, i.e., abnormal vs. normal. The sensitivity and specificity are calculated over all samples.

Evaluation results. An overview of all results is displayed in Tab. 7.1.

We evaluate both types of triplet loss for our proposed method with complexity and repertoire features as well as a combination of both. $Ours_{\text{seq}}$ denotes the triplet loss $trip_{seq}$, which is equal to [Ari+18b], and $Ours_{\text{GMA}}$ denotes our proposed triplet loss based on rating-similarity.

Ours_{seq} obtains poor sensitivity rates. We observe the best results for a classification of DA+MA vs. NS+NO at a sensitivity of 68% and a specificity of 70% with our complexity features. This is within the range of our baseline method results. For the medically more relevant discrimination between classes DA and not-DA, the results are far behind the baselines.

Ours_{GMA} shows an improved specificity ranging between 68% and 76%, which is among the best values of all evaluated methods. The sensitivity, however, is among

the worst results for DA vs. not-DA, lying between 39% and 56%. This is far from being useful in clinical practice. The classification of DA+MA vs. NS+NO achieves higher sensitivity of up to 64%, which is 15% less than the best baseline approach (B-1NN$_{16}$).

An inspection of the different features in *Ours* does not show an obvious trend. Ours$_{\text{seq}}$ (rep.) has the highest sensitivity among *Ours* for DA vs. not-DA, but drops by 23% for DA+MA vs. NS+NO, while the specificity shifts in the opposite direction. The combination of both features mostly does not change results significantly. The biggest difference is a *decrease* of sensitivity by 17% from Ours$_{\text{GMA}}$ (comp.) to Ours$_{\text{GMA}}$ (both).

Comparing *Ours*$_{\text{seq}}$ to *Ours*$_{\text{GMA}}$ reveals no considerable differences in results, except for the specificity being higher for *Ours*$_{\text{GMA}}$.

An analysis of the baseline methods shows diverse results. We evaluate **B-kNN** with word lengths of 16 and 32 frames, and $k \in \{1, 5\}$. The different parameters show relatively little influence on the results, with sensitivity in the range of 39-50% and specificity of 53-57% for classification DA vs. not-DA. When classifying DA+MA vs. NS+NP, the sensitivity improves to a maximum of 79%, and a specificity of around 70%.

The entropy baseline **B-ENT** achieves best results for a word length of 16 frames, at 83% sensitivity and 75% specificity for DA vs. not-DA, but decreases to 64% sensitivity and 66% specificity for DA+MA vs. NS+NO. For this classification task, the entropy over the full sequence obtains slightly better results at 74% sensitivity and a specificity of 70%.

The dynamic time warping repertoire baseline **B-DTW** with word length 16 achieves the same results as B-ENT$_{16}$, with slightly worse predictions for a word length of 32. Overall, results for baselines are slightly better when using a word length of 16 instead of 32.

The combination of baseline features does not lead to an improvement of results.

B-ENT is the only baseline that does not rely on our proposed complexity and repertoire features. Our results indicate that B-ENT correlates with GM quality, especially DA vs. not-DA, and that shorter words seem to achieve better results.

The other baselines rely on our proposed complexity (B-kNN) and repertoire (B-DTW) features (cf. Fig. 7.7). Their evaluation provides insights into the performance of the *features* using simple measures for similarity of motion words. The complexity feature better distinguishes between abnormal (DA+MA) and normal (NS+NO) movement quality compared to separating DA from not-DA. B-DTW obtains best results for classifying DA vs. not-DA, with sensitivity and specificity in the range of [Add+09; Add+10; Add+13; Gao+19]. This indicates that our proposed features capture characteristics of GMA. The foundation of both baselines is the identification of similar motion words, either by finding nearest neighbors in joint angle space, or by calculating the dynamic time warping distance. The target of our proposed

method is the development of a learned *representation of similarity* that is more relevant to GMA than the similarity measures used in the baseline methods.

We attribute the poor outcome of $Ours_{\text{seq}}$ to the focus on sequence-based clustering in the creation of the feature space, as explained previously (cf. Sec. 7.2.3), not to the proposed features.

$Ours_{\text{GMA}}$, on the other hand, shows a distribution of ratings in the feature space that is desired, i.e., embeddings are positioned according to their rating (cf. Fig. 7.6). Nevertheless, the results do not show the anticipated improvement.

7.2.5. Discussion

The most pressing question is: *Why did our approach not succeed?*

The answer is: We did not manage to create a feature space that projects input motion words into a neighborhood of similar motion words with similar ratings.

The initial triplet loss $trip_{seq}$ contains positives solely from the same sequence as the anchor, and all negatives are selected randomly from other sequences. This induces a grouping of words of the same sequence, as observed in Fig. 7.5. A connection across sequences is only established if the motion words have highly similar values and are thus projected to the same location in the feature space. In scenarios with clearly defined activities this works due to the repeated intra-class motion patterns. As shown in our experiments, it does not work for unconstrained infant motions.

Our adjustment of the initial triplet formation to $Ours_{\text{GMA}}$ leads to a smooth arrangement of training samples according to their labels (cf. Fig. 7.6). However, the information where to place an input sample with unknown label can only originate from the joint angles of the motion word. An appropriate placement is impossible if the CNN is trained to organize the words according to their *rating*, without sufficiently considering the actual *similarity* of motions.

With the current *labeling method*, many motion word labels will be incorrect due to the diversity inherent in spontaneous infant movements. One label for all motion words of a sequence is not enough, because these incorrectly labeled samples have the same influence on CNN training as correctly labeled samples. We believe that our approach can benefit from more *fine-grained* labeling. An ideal labeling consists of a large number of motions that are characteristic of a each GMA rating. This involves the laborious process of manual annotation and requires GMA experts.

The second question is: *Why is our approach still useful?*

Our baseline approaches show that the proposed complexity and repertoire features capture characteristics of general movements.

Training a CNN to embed features in a latent space with respect to specific properties has been successfully applied to motion sequences [Ari+18b]. The authors

demonstrate that the learned feature space better distinguishes fine-grained movements and creates a superior grouping of similar motions compared to DTW. If we can transfer these feature space properties to the domain of unconstrained infant movements our features will show their full potential.

In future work, we will target the collection of motion annotations by GMA experts to provide more precise labels for guiding the feature space creation.

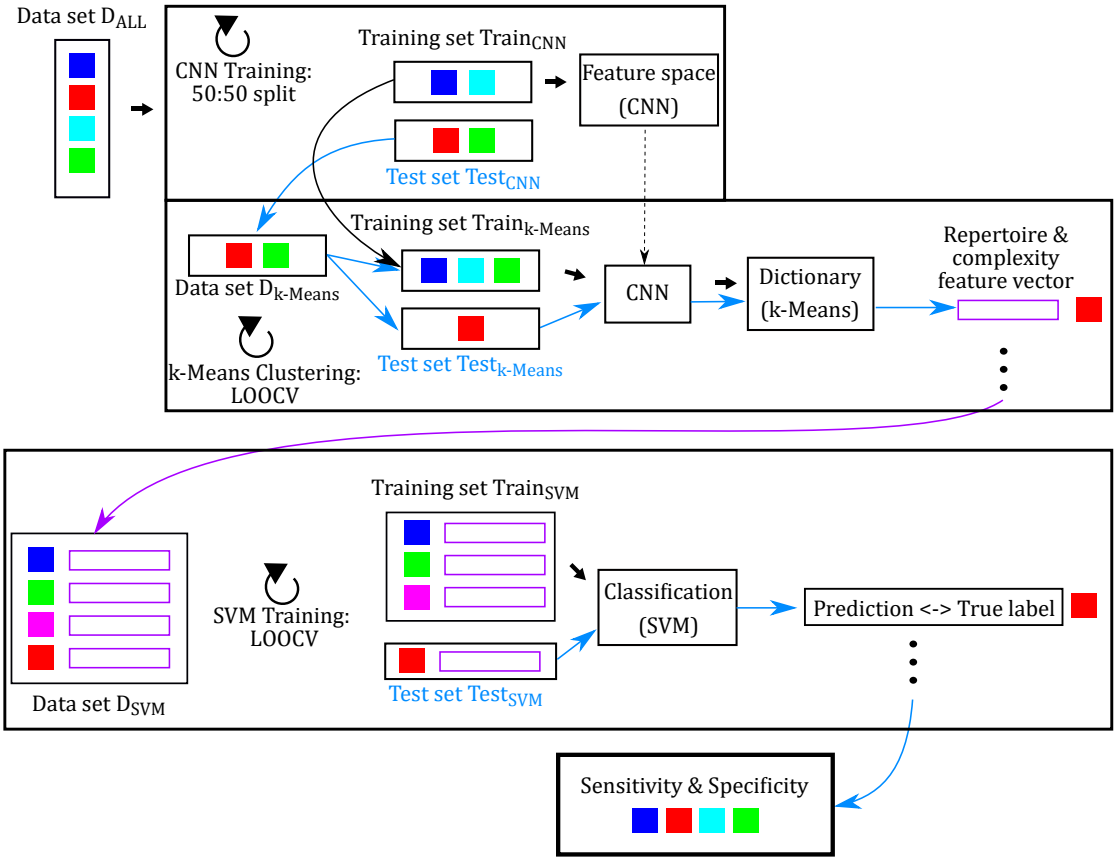


Figure 7.10.: Training procedure. For training the feature space, we split the training data D_{ALL} in two sets, and train the CNN with one set $Train_{CNN}$. To compute the repertoire and complexity features, we use a leave-one-out cross validation strategy (LOOCV) on the test set $Test_{CNN} == D_{k-Means}$, but *also* project motion words from $Train_{CNN}$ to the feature space to train a dictionary using k-Means. Then, we look up the motion words for the remaining sample in $Test_{k-Means}$ in the dictionary to create the feature vector. We repeat this until feature vectors for all samples in $D_{k-Means}$ are created. Then we switch $Train_{CNN}$ and $Test_{CNN}$ and repeat the training procedure until we have computed feature vectors for all samples in D_{ALL} to produce D_{SVM}. We apply LOOCV and train an SVM on $Train_{SVM}$, which we use to predict the label of the sample in $Test_{SVM}$. We accumulate the sensitivity and specificity for all samples.

8. Conclusion

In this final section of the thesis, we summarize our findings and discuss which future directions of research are opened up by our work.

8.1. Summary

In this thesis, we addressed the problem of automated infant motion analysis for early detection of cerebral palsy.

We extensively reviewed the literature on automated infant motion analysis, with particular focus on motion capture methods. This let us single out RGB-D sensors as the best choice as being the foundation of a low-cost, unintrusive, easy to use, and automated motion analysis system. Additionally, we found that the state of the art in infant motion capture is far behind the state of the art in general adult motion capture.

In our first effort to reduce this gap, we adapted a highly successful approach for pose estimation – used in the Kinect for XBox device – to infants and reduced the computational cost of the training procedure by a large margin. We identified limitations and proposed several extensions and showed how these contribute to a significant improvement of pose estimation accuracy.

However, we expect medical applications to require the highest precision possible, which is why we investigated more advanced methods for estimation of full body pose and shape instead of predicting only joint positions.

The accurate capture of *adult* shape and pose from RGB-D has been demonstrated by fitting a human body model to the data. For *infants*, no such body model existed. Opposed to standard practice in adult model creation, no set of high-quality 3D scans of infants in predefined poses is available. We introduced a method for learning an infant body model from low-quality, incomplete RGB-D data of freely moving infants. We publicly released the learned model to boost research on infant motion and shape analysis. We further provided a method for registering the model to the data for accurately capturing shape and pose of infants. Our experiments showed that this model carries enough information for letting experts perform accurate GMA ratings on synthetic model sequences of moving infants.

The literature review revealed a lack of evaluation of infant motion tracking approaches, which we believe is caused by the non-existence of a public infant data

set. We used our model to create a realistic synthetic RGB-D data set, which we made publicly available.

To complete the motion analysis pipeline, we first showed how human-interpretable measurements can help to distinguish between different diseases and provide quantitative measurements of infant motion. Finally, we used tracked poses for learning to predict the GMA class from pose sequences. We partition sequences into smaller motion words, and learn a feature space to group motion words that share characteristics. To classify a new sequence we analyze the statistics of the neighborhood of its motion words in feature space. Our proposed complexity and repertoire features displayed the capability of extracting relevant characteristics of general movements. The learned feature space, however, requires a better measure for motion word similarity. The formulation could be guided by manual annotations of motions that are representative of different GMA classes.

Our main contribution, the SMIL model, can be seen as a new tool to be used in many ways. One could imagine applications for medical purposes, ergonomics, or animation. It could be used not only for learning about impaired children, but lead to a better understanding of the development of motor skills by making movements measurable.

An automation of GMA allows widespread screening of children without the need for expensive hardware or setup effort. This could be especially useful in countries/areas with limited medical resources like equipment or trained experts. The described lightweight setup would allow medical workers to visit families at home, e.g., as is done in rural India for measuring children at risk of malnutrition [Sah+15].

8.2. Answers to Research Questions

In the beginning of this thesis, we have stated the following research questions.

Can we capture 3D infant motion in sufficient detail for automated motion analysis using a single RGB-D sensor?

We show that this is possible. Results for motion tracking using our learned model show roughly 99% correct poses in over 2 hours of moving infant recordings. Two experienced experts performed GMA on rendered SMIL sequences and had an agreement of roughly 90% and 80% with RGB ratings, with agreement being a rating difference $\leq$1 on a 1-10 scale.

Can we reliably infer GMA ratings from captured motions? We have showed that our proposed features can distinguish abnormal from normal movement quality to a certain extent. However, the sensitivity and specificity are not (yet) high enough to be relied upon in clinical practice. We are nevertheless convinced that our method can be adapted to create better representations of general movements, e.g. by the integration of GMA expert knowledge using manual annotations of relevant motions.

8.3. Future Work

Our proposed model-based method for capturing motion from RGB-D sequences can be extended to older children. Instead of the need to create an initial model for registering data of children, SMIL could be used. Different age groups to which the model generalizes well could be defined to successively train models for each age group based on the model from the previous one.

For capturing motions of older children, the method will need to handle poses that include turning, crawling, and walking. This challenge includes finding a trade-off between the camera being close to the subject to capture it in detail and the limited recording area of a camera in which a child can move before the camera loses track. A setup with multiple RGB-D sensors could solve this issue. However, there is a possibility of sensors disturbing each other, e.g., if projected patterns of multiple sensors overlap. Additionally, the advantage of having a simple, easy-to-use setup may be void if there is a need to calibrate a multi-camera setup prior to each use. Another possible solution is to use handheld cameras – these, however, may introduce motion blur and make it harder to fit a model to a point cloud that moves a lot within the camera coordinate system in successive frames.

A very active line of research for adult data is the estimate of 3D shape and pose from 2D RGB images. Many of the proposed approaches rely on the SMPL model, and since SMIL and SMPL share the same topology, these approaches should be easily transferable to infants. In the past, clinical institutions have recorded data sets for motion analysis like GMA using standard video cameras, since depth cameras were unavailable or unaffordable, or the possibilities of depth cameras were not realized. This data could be enhanced by lifting it to 3D to take full advantage of new methods for 3D motion analysis.

To make the model look more realistic, but also to capture more fine-grained details, an extension of SMIL to include facial expressions and finger movements is an obvious direction for future work. Making the interaction of infants/children with objects measurable may not only be relevant for medical assessment, but could also give new insights in the development of motor functions. In this context, the combination of RGB/RGB-D with other sensor types like haptic devices could be of interest.

Pressure sensor mats, which are often used for gait analysis, could complement vision-based systems. Learning to infer sensor measurements of one sensor type from the other, i.e., predicting foot pressure from body pose or vice versa, may increase the diagnostic value of each of the sensors.

Our proposed model makes it possible to quantify therapy or disease progression for diseases that affect motor abilities, e.g., spinal muscular atrophy (SMA). A universal parameter for distinguishing between normal and abnormal motor development for different age groups would be clinically relevant.

Besides motion, shape can be a marker of medical impairments. The body shape can provide information about malnutrition, e.g., in areas of the world where there is no widespread medical care. The lightweight system of an RGB-D capture system could enable medical workers to make reliable measurements at people's homes. This would allow to track shape development for identifying the best treatment on the one hand, and for monitoring the effect of therapy.

For the task of automated GMA, we aim to further improve results by applying a more fine-grained labeling of motion words instead of using one label for all words of a sequence. This requires GMA experts to manually identify motion words they consider to be relevant for the classification of a sequence, e.g., complex motions are representative of high motion quality.

The main contribution of this thesis is a new means of tracking infants, and we are excited to see how this is going to enable researchers to discover new applications and use-cases in the future.

Own Publications

[HPB+19] Nikolas Hesse, Sergi Pujades, Michael J. Black, Michael Arens, Ulrich G. Hofmann, and Sebastian Schroeder. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, https://doi.org/10.1109/TPAMI.2019.2917908, 2019.

[HBA+18] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Raphael Weinberger, and A. Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and RGB-D data set. In *Computer Vision – ECCV 2018 Workshops*, pages 32–49, 2018.

[HPR+18] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J. Black, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, Wolfgang Müller-Felber, and A. Sebastian Schroeder. 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*, pages 792–800, 2018.

[HSMF+17] Nikolas Hesse, A. Sebastian Schröder, Wolfgang Müller-Felber, Christoph Bodensteiner, Michael Arens, and Ulrich G. Hofmann. Body pose estimation in depth images for infant motion analysis. In *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1909–1912, 2017.

[HSSMF+17] Nikolas Hesse, A. Sebastian Schroeder, Wolfgang Müller-Felber, Christoph Bodensteiner, Michael Arens, and Ulrich G. Hofmann. Markerless motion analysis for early detection of infantile movement disorders. In *EMBEC & NBC 2017*, pages 197–200, 2017.

[TWBH+17] U. Tacke, H. Weigand-Brunnhölzl, A. Hilgendorff, R. M. Giese, A. W. Flemmer, H. König, B. Warken-Madelung, M. Arens, N. Hesse, and A. S. Schroeder. Entwicklungsneurologie – vernetzte medizin und neue perspektiven. *Der Nervenarzt*, pages 1395–1401, 2017.

[HSBA15] Nikolas Hesse, Gregor Stachowiak, Timo Breuer, and Michael Arens. Estimating body pose of infants in depth images using random ferns. In *IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 427–435, 2015.

List of Figures

List of Tables

Bibliography

[AB15] I. Akhter and M. J. Black. "Pose-conditioned joint angle limits for 3D human pose reconstruction". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1446–1455. DOI: 10.1109/CVPR.2015.7298751.

[Ach+16] F. Achilles et al. "Patient MoCap: Human Pose Estimation Under Blanket Occlusion for Hospital Monitoring Applications". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*. Ed. by S. Ourselin et al. Cham: Springer International Publishing, 2016, pp. 491–499. ISBN: 978-3-319-46720-7.

[Add+09] L. Adde et al. "Using computer-based video analysis in the study of fidgety movements". In: *Early Human Development* 85.9 (2009), pp. 541 –547. ISSN: 0378-3782. DOI: https://doi.org/10.1016/j.earlhumdev.2009.05.003.

[Add+10] L. Adde et al. "Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study". In: *Developmental Medicine & Child Neurology* 52.8 (2010), pp. 773–778.

[Add+13] L. Adde et al. "Identification of fidgety movements and prediction of CP by the use of computer-based video analysis is more accurate when based on two video recordings". In: *Physiotherapy Theory and Practice* 29.6 (2013), pp. 469–475. DOI: 10.3109/09593985.2012.757404.

[Add+18] L. Adde et al. "Characteristics of general movements in preterm infants assessed by computer-based video analysis". In: *Physiotherapy Theory and Practice* 34.4 (2018), pp. 286–292. DOI: 10.1080/09593985.2017.1391908.

[All+03] B. Allen et al. "The Space of Human Body Shapes: Reconstruction and Parameterization from Range Scans". In: *ACM Trans. Graph.* 22.3 (July 2003), pp. 587–594. ISSN: 0730-0301. DOI: 10.1145/882262.882311.

[All+17] T. Alldieck et al. "Optical Flow-Based 3D Human Motion Estimation from Monocular Video". In: *Pattern Recognition*. Ed. by V. Roth and T. Vetter. Cham: Springer International Publishing, 2017, pp. 347–360. ISBN: 978-3-319-66709-6.

[All+18] T. Alldieck et al. “Video Based Reconstruction of 3D People Models”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8387–8397. DOI: 10.1109/CVPR.2018.00875.

[And+14] M. Andriluka et al. “2D Human Pose Estimation: New Benchmark and State of the Art Analysis”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3686–3693. DOI: 10.1109/CVPR.2014.471.

[And+18] M. Andriluka et al. “PoseTrack: A Benchmark for Human Pose Estimation and Tracking”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5167–5176. DOI: 10.1109/CVPR.2018.00542.

[Ang+05] D. Anguelov et al. “SCAPE: Shape Completion and Animation of People”. In: *ACM Trans. Graph.* 24.3 (July 2005), pp. 408–416. ISSN: 0730-0301. DOI: 10.1145/1073204.1073207.

[Ari+18a] A. Aristidou et al. “Self-similarity Analysis for Motion Capture Cleaning”. In: vol. 37. 2. 2018, pp. 297–309. DOI: 10.1111/cgf.13362.

[Ari+18b] A. Aristidou et al. “Deep Motifs and Motion Signatures”. In: SIGGRAPH Asia ’18 (2018), 187:1–187:13. DOI: 10.1145/3272127.3275038.

[Baa+11] A. Baak et al. “A data-driven approach for real-time full body pose reconstruction from a depth camera”. In: *2011 International Conference on Computer Vision*. 2011, pp. 1092–1099. DOI: 10.1109/ICCV.2011.6126356.

[Baa+13] A. Baak et al. “A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera”. In: *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. Ed. by A. Fossati et al. London: Springer London, 2013, pp. 71–98. ISBN: 978-1-4471-4640-7. DOI: 10.1007/978-1-4471-4640-7_5.

[Bam+18] C. S. Bamji et al. “IMpixel 65nm BSI 320MHz demodulated TOF Image sensor with 3μm global shutter pixels and analog binning”. In: *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*. 2018, pp. 94–96. DOI: 10.1109/ISSCC.2018.8310200.

[Ber+11] I. Bernhardt et al. “Inter- and intra-observer agreement of Prechtl’s method on the qualitative assessment of general movements in preterm, term and young infants”. In: *Early Human Development* 87.9 (2011), pp. 633 –639. ISSN: 0378-3782. DOI: https://doi.org/10.1016/j.earlhumdev.2011.04.017.

[Bes89] P. J. Besl. “Active optical range imaging sensors”. In: *Advances in machine vision*. Springer, 1989, pp. 1–63.

[BHHA05] C. H. Blauw-Hospers and M. Hadders-Algra. "A systematic review of the effects of early intervention on motor development". In: *Developmental Medicine & Child Neurology* 47.6 (2005), pp. 421–432. DOI: 10.1017/S0012162205000824.

[Ble15] Blender. *Free and open 3D creation software*. www.blender.org. Sept. 2015.

[Bog+14] F. Bogo et al. "FAUST: Dataset and Evaluation for 3D Mesh Registration". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3794–3801. DOI: 10.1109/CVPR.2014.491.

[Bog+15] F. Bogo et al. "Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2300–2308. DOI: 10.1109/ICCV.2015.265.

[Bog+16] F. Bogo et al. "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image". In: *Computer Vision – ECCV 2016*. Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016.

[Bog+17] F. Bogo et al. "Dynamic FAUST: Registering Human Bodies in Motion". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5573–5582. DOI: 10.1109/CVPR.2017.591.

[Bos+13] M. Bosanquet et al. "A systematic review of tests to predict cerebral palsy in young children". In: *Developmental Medicine & Child Neurology* 55.5 (2013), pp. 418–426.

[Bou+10] H. Bouwstra et al. "Predictive value of definitely abnormal general movements in the general population". In: *Developmental Medicine & Child Neurology* 52.5 (2010), pp. 456–461. DOI: 10.1111/j.1469-8749.2009.03529.x.

[Boy+11] C. A. Boyle et al. "Trends in the Prevalence of Developmental Disabilities in US Children, 1997–2008". In: *Pediatrics* 127.6 (2011), pp. 1034–1042. ISSN: 0031-4005. DOI: 10.1542/peds.2010-2989.

[Bud+11] M. Budiu et al. "Parallelizing the training of the Kinect body parts labeling algorithm". In: *Big Learning: Algorithms, Systems and Tools for Learning at Scale* (2011), pp. 1–6.

[Buy+14] K. Buys et al. "An adaptable system for RGB-D based human body detection and pose estimation". In: *Journal of Visual Communication and Image Representation* 25.1 (2014). Visual Understanding and Applications with RGB-D Cameras, pp. 39 –52. ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2013.03.011.

[BV99] V. Blanz and T. Vetter. "A Morphable Model for the Synthesis of 3D Faces". In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194. ISBN: 0-201-48560-5. DOI: 10.1145/311535.311556.

[Cao+17] Z. Cao et al. "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1302–1310. DOI: 10.1109/CVPR.2017.143.

[Car+16] J. Carreira et al. "Human Pose Estimation with Iterative Error Feedback". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4733–4742. DOI: 10.1109/CVPR.2016.512.

[Cen+17] A. Cenci et al. "Movements Analysis of Preterm Infants by Using Depth Sensor". In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*. IML '17. Liverpool, United Kingdom: ACM, 2017, 12:1–12:9. ISBN: 978-1-4503-5243-7. DOI: 10.1145/3109761.3109773.

[CF13] T. J. Cashman and A. W. Fitzgibbon. "What Shape Are Dolphins? Building 3D Morphable Models from 2D Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 232–244. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.68.

[Cha+11] J. C. P. Chan et al. "A Virtual Reality Dance Training System Using Motion Capture Technology". In: *IEEE Transactions on Learning Technologies* 4.2 (2011), pp. 187–195. ISSN: 1939-1382. DOI: 10.1109/TLT.2010.27.

[Che+13] Y. Chen et al. "Tensor-Based Human Body Modeling". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 105–112. DOI: 10.1109/CVPR.2013.21.

[Che97] P. D. Cheney. "Pathophysiology of the corticospinal system and basal ganglia in cerebral palsy". In: *Mental Retardation and Developmental Disabilities Research Reviews* 3.2 (1997), pp. 153–167. DOI: 10.1002/(SICI)1098-2779(1997)3:2<153::AID-MRDD7>3.0.CO;2-S.

[Chr+14] D. Christensen et al. "Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning–Autism and Developmental Disabilities Monitoring Network, USA, 2008". In: *Developmental Medicine & Child Neurology* 56.1 (2014), pp. 59–65.

[Cla+12] R. A. Clark et al. "Validity of the Microsoft Kinect for assessment of postural control". In: *Gait & Posture* 36.3 (2012), pp. 372 –377. ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2012.03.033.

[Cos+03] M. Costa et al. “Multiscale entropy analysis of human gait dynamics”. In: *Physica A: Statistical Mechanics and its Applications* 330.1 (2003). RANDOMNESS AND COMPLEXITY: Proceedings of the International Workshop in honor of Shlomo Havlin’s 60th birthday, pp. 53 – 60. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2003.08.022.

[Cui+13] Y. Cui et al. “KinectAvatar: Fully Automatic Body Capture Using a Single Kinect”. In: *Computer Vision - ACCV 2012 Workshops*. Ed. by J.-I. Park and J. Kim. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 133–147. ISBN: 978-3-642-37484-5.

[DeC+98] D. DeCarlo et al. “An Anthropometric Face Model Using Variational Techniques”. In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’98. New York, NY, USA: ACM, 1998, pp. 67–74. ISBN: 0-89791-999-8. DOI: 10.1145/280814.280823.

[Dir+11] T. Dirks et al. “Differences Between the Family-Centered “COPCA” Program and Traditional Infant Physical Therapy Based on Neurodevelopmental Treatment Principles”. In: *Physical Therapy* 91.9 (Sept. 2011), pp. 1303–1322. ISSN: 0031-9023. DOI: 10.2522/ptj.20100207.

[DK+12] C. Disselhorst-Klug et al. “Introduction of a method for quantitative evaluation of spontaneous motor activity development with age in infants”. In: *Experimental Brain Research* 218.2 (2012), pp. 305–313. ISSN: 1432-1106. DOI: 10.1007/s00221-012-3015-x.

[Ein+97] C. Einspieler et al. “The qualitative assessment of general movements in preterm, term and young infants – review of the methodology”. In: *Early Human Development* 50.1 (1997), pp. 47 –60. ISSN: 0378-3782. DOI: https://doi.org/10.1016/S0378-3782(97)00092-3.

[Elh+15] A. Elhayek et al. “Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3810–3818. DOI: 10.1109/CVPR.2015.7299005.

[Elh+17] A. Elhayek et al. “MARCOnI–ConvNet-Based MARker-Less Motion Capture in Outdoor and Indoor Scenes”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.3 (2017), pp. 501–514. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2557779.

[EP05] C. Einspieler and H. F. R. Prechtl. “Prechtl’s assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system”. In: *Mental Retardation and Developmental Disabilities Research Reviews* 11.1 (2005), pp. 61–67. DOI: 10.1002/mrdd.20051.

[Fan+12] M. Fan et al. “Augmenting Gesture Recognition with Erlang-Cox Models to Identify Neurological Disorders in Premature Babies”. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp ’12. Pittsburgh, Pennsylvania: ACM, 2012, pp. 411–420. ISBN: 978-1-4503-1224-0. DOI: 10.1145/2370216.2370278.

[Fan+15] X. Fan et al. “Combining local appearance and holistic view: Dual-Source Deep Neural Networks for human pose estimation”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1347–1355. DOI: 10.1109/CVPR.2015.7298740.

[FB12] O. Freifeld and M. J. Black. “Lie Bodies: A Manifold Representation of 3D Human Shape”. In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–14. ISBN: 978-3-642-33718-5.

[FB81] M. A. Fischler and R. C. Bolles. “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”. In: *Commun. ACM* 24.6 (June 1981), pp. 381–395.

[Fis+99] J. S. Fischer et al. “The Multiple Sclerosis Functional Composite measure (MSFC): an integrated approach to MS clinical outcome assessment”. In: *Multiple Sclerosis Journal* 5.4 (1999). PMID: 10467383, pp. 244–250. DOI: 10.1177/135245859900500409.

[Gan+12] V. Ganapathi et al. “Real-Time Human Pose Tracking from Range Data”. In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 738–751. ISBN: 978-3-642-33783-3.

[Gao+19] Y. Gao et al. *Towards Reliable, Automated General Movement Assessment for Perinatal Stroke Screening in Infants Using Wearable Accelerometers*. 2019.

[Gen11] J. Geng. “Structured-light 3D surface imaging: a tutorial”. In: *Advances in Optics and Photonics* 3.2 (2011), pp. 128–160.

[Gir+11] R. Girshick et al. “Efficient regression of general-activity human poses from depth images”. In: *2011 International Conference on Computer Vision*. 2011, pp. 415–422. DOI: 10.1109/ICCV.2011.6126270.

[Gor+09] J. W. Gorter et al. “Use of the GMFCS in infants with CP: the need for reclassification at age 2 years or older”. In: *Developmental Medicine & Child Neurology* 51.1 (2009), pp. 46–52.

[Gra+12] D Gravem et al. “Assessment of infant movement with a compact wireless accelerometer system”. In: *Journal of Medical Devices* 6.2 (2012), p. 021013.

[GS01] R. Gross and J. Shi. “The CMU motion of body (mobo) database”. In: (2001).

[GS03] H. K. Graham and P. Selber. “Muscoskeletal Aspects of Cerebral Palsy”. In: *The Journal of Bone and Joint Surgery. British volume* 85-B.2 (2003), pp. 157–166. DOI: 10.1302/0301-620X.85B2.14066.

[Gua+09] P. Guan et al. “Estimating human shape and pose from a single image”. In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 1381–1388. DOI: 10.1109/ICCV.2009.5459300.

[HA+04] M. Hadders-Algra et al. “Quality of general movements and the development of minor neurological dysfunction at toddler and school age”. In: *Clinical Rehabilitation* 18.3 (2004), pp. 287–299.

[HA+17] M. Hadders-Algra et al. “Effect of early intervention in infants at very high risk of cerebral palsy: a systematic review”. In: *Developmental Medicine & Child Neurology* 59.3 (2017), pp. 246–258.

[HA04] M. Hadders-Algra. “General movements: a window for early identification of children at high risk for developmental disorders”. In: *The Journal of Pediatrics* 145.2, Supplement (2004). Cerebral palsy: Selected conference papers, S12 –S18. ISSN: 0022-3476. DOI: https://doi.org/10.1016/j.jpeds.2004.05.017.

[HA10] M. Hadders-Algra. “Variation and variability: key words in human motor development”. In: *Physical therapy* 90.12 (2010), pp. 1823–1837.

[HA14] M. Hadders-Algra. “Early Diagnosis and Early Intervention in Cerebral Palsy”. In: *Frontiers in Neurology* 5 (2014), p. 185. ISSN: 1664-2295. DOI: 10.3389/fneur.2014.00185.

[HAP18] M. Hadders-Algra and H. Philippi. “Predictive validity of the General Movements Assessment: type of population versus type of assessment”. In: *Developmental Medicine & Child Neurology* 60.11 (2018), pp. 1186–1186. DOI: 10.1111/dmcn.14000.

[Haq+16] A. Haque et al. “Towards Viewpoint Invariant 3D Human Pose Estimation”. In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe et al. Cham: Springer International Publishing, 2016, pp. 160–177. ISBN: 978-3-319-46448-0.

[Has+09] N. Hasler et al. “A Statistical Model of Human Pose and Body Shape”. In: *Computer Graphics Forum* 28.2 (2009), pp. 337–346. DOI: 10.1111/j.1467-8659.2009.01373.x.

[Has+10] N. Hasler et al. “Multilinear pose and body shape estimation of dressed subjects from image sets”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 1823–1830. DOI: 10.1109/CVPR.2010.5539853.

[He+17] K. He et al. “Mask R-CNN”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[Hei+10] F. Heinze et al. “Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy”. In: *Medical & Biological Engineering & Computing* 48.8 (2010), pp. 765–772. ISSN: 1741-0444. DOI: 10.1007/s11517-010-0624-z.

[Hel+13a] T. Helten et al. “Personalization and Evaluation of a Real-Time Depth-Based Full Body Tracker”. In: *2013 International Conference on 3D Vision - 3DV 2013*. 2013, pp. 279–286. DOI: 10.1109/3DV.2013.44.

[Hel+13b] T. Helten et al. “Real-Time Body Tracking with One Depth Camera and Inertial Sensors”. In: *2013 IEEE International Conference on Computer Vision*. 2013, pp. 1105–1112. DOI: 10.1109/ICCV.2013.141.

[Hen+07] S. Henderson et al. *Movement assessment battery for children–second edition*. 2007.

[Her+15] A. Herskind et al. “Early identification and intervention in cerebral palsy”. In: *Developmental Medicine & Child Neurology* 57.1 (2015), pp. 29–36. DOI: 10.1111/dmcn.12531.

[Hes+15] N. Hesse et al. “Estimating Body Pose of Infants in Depth Images Using Random Ferns”. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2015, pp. 427–435. DOI: 10.1109/ICCVW.2015.63.

[Hes+17a] N. Hesse et al. “Body pose estimation in depth images for infant motion analysis”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017, pp. 1909–1912. DOI: 10.1109/EMBC.2017.8037221.

[Hes+17b] N. Hesse et al. “Markerless Motion Analysis for Early Detection of Infantile Movement Disorders”. In: *EMBEC & NBC 2017: Joint Conference of the European Medical and Biological Engineering Conference (EMBEC) and the Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC), Tampere, Finland, June 2017*. Ed. by H. Eskola et al. Singapore: Springer Singapore, 2017, pp. 197–200. ISBN: 978-981-10-5122-7. DOI: 10.1007/978-981-10-5122-7_50.

[Hes+18] 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*. Ed. by A. F. Frangi et al. Cham: Springer International Publishing, 2018, pp. 792–800. ISBN: 978-3-030-00928-1.

[Hes+19a] N. Hesse et al. “Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set”. In: *Computer Vision – ECCV 2018 Workshops*. Ed. by L. Leal-Taixé and S. Roth. Cham: Springer International Publishing, 2019, pp. 32–49. ISBN: 978-3-030-11024-6.

[Hes+19b] N. Hesse et al. “Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D sequences”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019). ISSN: 1939-3539.

[Hie+11] T. Hielkema et al. “Does physiotherapeutic intervention affect motor outcome in high-risk infants? An approach combining a randomized controlled trial and process evaluation”. In: *Developmental Medicine & Child Neurology* 53.3 (2011), e8–e15.

[Him13] K. Himmelmann. “Chapter 15 - Epidemiology of cerebral palsy”. In: *Pediatric Neurology Part I*. Ed. by O. Dulac et al. Vol. 111. Handbook of Clinical Neurology. Elsevier, 2013, pp. 163 –167. DOI: https://doi.org/10.1016/B978-0-444-52891-9.00015-4.

[Hir+11] D. A. Hirshberg et al. “Evaluating the automated alignment of 3D human body scans”. In: *2nd Int. Conf. 3D Body Scanning Technologies*. 2011, pp. 76–86.

[Hir+12] D. A. Hirshberg et al. “Coregistration: Simultaneous Alignment and Modeling of Articulated 3D Shape”. In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 242–255. ISBN: 978-3-642-33783-3.

[Hua+13] C. Huang et al. “Robust Human Body Shape and Pose Tracking”. In: *2013 International Conference on 3D Vision - 3DV 2013*. 2013, pp. 287–294. DOI: 10.1109/3DV.2013.45.

[Hua+17] Y. Huang et al. “Towards Accurate Marker-Less Human Shape and Pose Estimation over Time”. In: *2017 International Conference on 3D Vision (3DV)*. 2017, pp. 421–430. DOI: 10.1109/3DV.2017.00055.

[Hua+18] C. P. Huang et al. “Tracking-by-Detection of 3D Human Shapes: From Surfaces to Volumes”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.8 (2018), pp. 1994–2008. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2017.2740308.

[Hut06] J. L. Hutton. “Cerebral Palsy Life Expectancy”. In: *Clinics in Perinatology* 33.2 (2006). Perinatal Causes of Cerebral Palsy, pp. 545 –555. ISSN: 0095-5108. DOI: https://doi.org/10.1016/j.clp.2006.03.016.

[Iar+14] S. Iarlori et al. “RGBD camera monitoring system for Alzheimer’s disease assessment using Recurrent Neural Networks with Parametric Bias action recognition”. In: *IFAC Proceedings Volumes* 47.3 (2014). 19th IFAC World Congress, pp. 3863 –3868. ISSN: 1474-6670. DOI: https://doi.org/10.3182/20140824-6-ZA-1003.02199.

[Ins+16] E. Insafutdinov et al. “DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model”. In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe et al. Cham: Springer International Publishing, 2016, pp. 34–50. ISBN: 978-3-319-46466-4.

[Ins+17] E. Insafutdinov et al. “ArtTrack: Articulated Multi-Person Tracking in the Wild”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1293–1301. DOI: 10.1109/CVPR.2017.142.

[Jen+07] K. M. Jenks et al. “The Effect of Cerebral Palsy on Arithmetic Accuracy is Mediated by Working Memory, Intelligence, Early Numeracy, and Instruction Time”. In: *Developmental Neuropsychology* 32.3 (2007), pp. 861–879. DOI: 10.1080/87565640701539758.

[Jia+18] C. Jiang et al. “Determining if wearable sensors affect infant leg movement frequency”. In: *Developmental Neurorehabilitation* 21.2 (2018), pp. 133–136. DOI: 10.1080/17518423.2017.1331471.

[Jun+15] H. Y. Jung et al. “Random tree walk toward instantaneous 3D human pose estimation”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 2467–2474. DOI: 10.1109/CVPR.2015.7298861.

[Jun+16] H. Y. Jung et al. “A Sequential Approach to 3D Human Pose Estimation: Separation of Localization and Identification of Body Joints”. In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe et al. Cham: Springer International Publishing, 2016, pp. 747–761. ISBN: 978-3-319-46454-1.

[Kan+13] N. Kanemaru et al. “Specific characteristics of spontaneous movements in preterm infants at term age are associated with developmental delays at age 3 years”. In: *Developmental Medicine & Child Neurology* 55.8 (2013), pp. 713–721.

[Kan+14] N. Kanemaru et al. “Jerky spontaneous movements at term age in preterm infants who later developed cerebral palsy”. In: *Early Human Development* 90.8 (2014), pp. 387 –392. ISSN: 0378-3782. DOI: http://dx.doi.org/10.1016/j.earlhumdev.2014.05.004.

[Kan+16] A. Kanazawa et al. “Learning 3D Deformation of Animals from 2D Images”. In: *Computer Graphics Forum* 35.2 (2016), pp. 365–374. ISSN: 1467-8659. DOI: 10.1111/cgf.12838.

[Kan+18a] A. Kanazawa et al. “End-to-End Recovery of Human Shape and Pose”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7122–7131. DOI: 10.1109/CVPR.2018.00744.

[Kan+18b] A. Kanazawa et al. “Learning Category-Specific Mesh Reconstruction from Image Collections”. In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari et al. Cham: Springer International Publishing, 2018, pp. 386–402. ISBN: 978-3-030-01267-0.

[Kar+08] D. Karch et al. “Quantification of the segmental kinematics of spontaneous infant movements”. In: *Journal of Biomechanics* 41.13 (2008), pp. 2860 –2867. ISSN: 0021-9290. DOI: https://doi.org/10.1016/j.jbiomech.2008.06.033.

[Kar+10] D. Karch et al. “Quantitative score for the evaluation of kinematic recordings in neuropediatric diagnostics. Detection of complex patterns in spontaneous limb movements”. In: *Methods of information in medicine* 49.5 (2010), pp. 526–530. ISSN: 0026-1270. DOI: 10.3414/me09-02-0034.

[Kar+12] D. Karch et al. “Kinematic assessment of stereotypy in spontaneous movements in infants”. In: *Gait & Posture* 36.2 (2012), pp. 307 –311. ISSN: 0966-6362. DOI: https://doi.org/10.1016/j.gaitpost.2012.03.017.

[Kar11] D. Karch. “Quantitative Analyse der Spontanmotorik von Säuglingen für die Prognose der infantilen Cerebralparese”. In: (2011).

[Kes+17] L. Keselman et al. “Intel(R) RealSense(TM) Stereoscopic Depth Cameras”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, pp. 1267–1276. DOI: 10.1109/CVPRW.2017.167.

[Kha+18] M. H. Khan et al. “Detection of Infantile Movement Disorders in Video Data Using Deformable Part-Based Model”. In: *Sensors* 18.10 (2018). ISSN: 1424-8220. DOI: 10.3390/s18103202.

[Kim+17] M. Kim et al. “Data-driven Physics for Human Soft Tissue Animation”. In: *ACM Trans. Graph.* 36.4 (July 2017), 54:1–54:12. ISSN: 0730-0301. DOI: 10.1145/3072959.3073685.

[Kin09] D. E. King. “Dlib-ml: A Machine Learning Toolkit”. In: *Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.

[Kin15] D. E. King. “Max-margin object detection”. In: *arXiv preprint arXiv: 1502.00046* (2015).

[Kno+06] S. Knoop et al. “Sensor fusion for 3D human body tracking with an articulated 3D body model”. In: *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* 2006, pp. 1686–1691. DOI: 10.1109/ROBOT.2006.1641949.

[Kol+13] B. Kolb et al. “Chapter 2 - Brain Plasticity in the Developing Brain”. In: *Changing Brains*. Ed. by M. M. Merzenich et al. Vol. 207. Progress in Brain Research. Elsevier, 2013, pp. 35 –64. DOI: https://doi.org/10.1016/B978-0-444-63327-9.00005-9.

[Kon+14] P. Kontschieder et al. “Quantifying Progression of Multiple Sclerosis via Classification of Depth Videos”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Ed. by P. Golland et al. Cham: Springer International Publishing, 2014, pp. 429–437. ISBN: 978-3-319-10470-6.

[Kow+18] A. Kowdle et al. “The Need 4 Speed in Real-time Dense Visual Tracking”. In: *SIGGRAPH Asia 2018 Technical Papers*. SIGGRAPH Asia ’18. Tokyo, Japan: ACM, 2018, 220:1–220:14. ISBN: 978-1-4503-6008-1. DOI: 10.1145/3272127.3275062.

[KS14] V. Kazemi and J. Sullivan. “One millisecond face alignment with an ensemble of regression trees”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1867–1874. DOI: 10.1109/CVPR.2014.241.

[Kuh+17] A. Kuhner et al. “An online system for tracking the performance of Parkinson’s patients”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 1664–1669. DOI: 10.1109/IROS.2017.8205977.

[Kur83] J. F. Kurtzke. “Rating neurologic impairment in multiple sclerosis”. In: *Neurology* 33.11 (1983), pp. 1444–1444. ISSN: 0028-3878. DOI: 10.1212/WNL.33.11.1444.

[Kwo+18] A. K. Kwong et al. “Predictive validity of spontaneous early infant movement for later cerebral palsy: a systematic review”. In: *Developmental Medicine & Child Neurology* (2018).

[Lal+13] J. Lallemand et al. “Multi-task Forest for Human Pose Estimation in Depth Images”. In: *2013 International Conference on 3D Vision - 3DV 2013*. 2013, pp. 271–278. DOI: 10.1109/3DV.2013.43.

[Las+17] C. Lassner et al. “Unite the People: Closing the Loop Between 3D and 2D Human Representations”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4704–4713. DOI: 10.1109/CVPR.2017.500.

[LB14] M. M. Loper and M. J. Black. “OpenDR: An Approximate Differentiable Renderer”. In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet et al. Cham: Springer International Publishing, 2014, pp. 154–169. ISBN: 978-3-319-10584-0.

[LC85] H.-J. Lee and Z. Chen. “Determination of 3D human body postures from a single view”. In: *Computer Vision, Graphics, and Image Processing* 30.2 (1985), pp. 148 –168. ISSN: 0734-189X. DOI: https://doi.org/10.1016/0734-189X(85)90094-5.

[Lee+02] J. Lee et al. “Interactive Control of Avatars Animated with Human Motion Data”. In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’02. San Antonio, Texas: ACM, 2002, pp. 491–500. ISBN: 1-58113-521-1. DOI: 10.1145/566570.566607.

[Lef+13] D. Lefloch et al. “Technical Foundation and Calibration Methods for Time-of-Flight Cameras”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications: Dagstuhl 2012 Seminar on Time-of-Flight Imaging and GCPR 2013 Workshop on Imaging New Modalities*. Ed. by M. Grzegorzek et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 3–24. ISBN: 978-3-642-44964-2. DOI: 10.1007/978-3-642-44964-2_1.

[Li+17] T. Li et al. “Learning a Model of Facial Shape and Expression from 4D Scans”. In: *ACM Trans. Graph.* 36.6 (Nov. 2017), 194:1–194:17. ISSN: 0730-0301. DOI: 10.1145/3130800.3130813.

[Liu+13] Y. Liu et al. “Markerless Motion Capture of Multiple Characters Using Multiview Image Segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.11 (2013), pp. 2720–2735. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.47.

[Lop] M. Loper. *Chumpy*. http://chumpy.org.

[Lop+14] M. Loper et al. “MoSh: Motion and Shape Capture from Sparse Markers”. In: *ACM Trans. Graph.* 33.6 (Nov. 2014), 220:1–220:13. ISSN: 0730-0301. DOI: 10.1145/2661229.2661273.

[Lop+15] M. Loper et al. “SMPL: A Skinned Multi-person Linear Model”. In: *ACM Trans. Graph.* 34.6 (Oct. 2015), 248:1–248:16. ISSN: 0730-0301. DOI: 10.1145/2816795.2818013.

[MÖ7] M. Müller. *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer-Verlag, 2007. ISBN: 3540740473.

[MA07] S. Mitra and T. Acharya. “Gesture Recognition: A Survey”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.3 (2007), pp. 311–324. ISSN: 1094-6977. DOI: 10.1109/TSMCC.2007.893280.

[Mac+17] A. Machireddy et al. “A video/IMU hybrid system for movement estimation in infants”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017, pp. 730–733. DOI: 10.1109/EMBC.2017.8036928.

[Mak] MakeHuman. *Open source tool for making 3D characters*. URL: www.makehumancommunity.org.

[Mar+16] T. von Marcard et al. “Human Pose Estimation from Video and IMUs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2016), pp. 1533–1547. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2522398.

[Mar+17a] P. B. Marschik et al. “A Novel Way to Measure and Predict Development: A Heuristic Approach to Facilitate the Early Detection of Neurodevelopmental Disorders”. In: *Current Neurology and Neuroscience Reports* 17.5 (2017), p. 43. ISSN: 1534-6293. DOI: 10.1007/s11910-017-0748-8.

[Mar+17b] J. Martinez et al. “A Simple Yet Effective Baseline for 3D Human Pose Estimation”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2659–2668. DOI: 10.1109/ICCV.2017.288.

[Mar+18] T. von Marcard et al. “Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera”. In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari et al. Cham: Springer International Publishing, 2018, pp. 614–631. ISBN: 978-3-030-01249-6.

[McI+11] S. McIntyre et al. “Cerebral palsy - don’t delay”. In: *Developmental disabilities research reviews* 17.2 (2011), pp. 114–129.

[Meh+17a] D. Mehta et al. “Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision”. In: *2017 International Conference on 3D Vision (3DV)*. 2017, pp. 506–516. DOI: 10.1109/3DV.2017.00064.

[Meh+17b] D. Mehta et al. “VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera”. In: *ACM Trans. Graph.* 36.4 (July 2017), 44:1–44:14. ISSN: 0730-0301. DOI: 10.1145/3072959.3073596.

[Mei+06] L. Meinecke et al. “Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy”. In: *Human Movement Science* 25.2 (2006), pp. 125 –144. ISSN: 0167-9457. DOI: https://doi.org/10.1016/j.humov.2005.09.012.

[MH08] L. v. d. Maaten and G. Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[MJ+18] M. J. Marín-Jiménez et al. “3D human pose estimation from depth maps using a deep combination of poses”. In: *Journal of Visual Communication and Image Representation* 55 (2018), pp. 627 –639. ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2018.07.010.

[MM14] P. Morgan and J. McGinley. “Gait function and decline in adults with cerebral palsy: a systematic review”. In: *Disability and Rehabilitation* 36.1 (2014). PMID: 23594053, pp. 1–9. DOI: 10.3109/09638288.2013.775359.

[MN17] F. Moreno-Noguer. “3D Human Pose Estimation from a Single Image via Distance Matrix Regression”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1561–1570. DOI: 10.1109/CVPR.2017.170.

[Moe+06] T. B. Moeslund et al. “A survey of advances in vision-based human motion capture and analysis”. In: *Computer Vision and Image Understanding* 104.2 (2006), pp. 90 –126. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2006.08.002.

[Moo+18] G. Moon et al. “V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5079–5088. DOI: 10.1109/CVPR.2018.00533.

[Mor+16a] C. Morgan et al. “Effectiveness of motor interventions in infants with cerebral palsy: a systematic review”. In: *Developmental Medicine & Child Neurology* 58.9 (2016), pp. 900–909.

[Mor+16b] C. Morrison et al. “Vision-based body tracking: turning Kinect into a clinical tool”. In: *Disability and Rehabilitation: Assistive Technology* 11.6 (2016), pp. 516–520. DOI: 10.3109/17483107.2014.989419.

[MS13] M. Martinez and R. Stiefelhagen. “Kinect Unleashed: Getting Control over High Resolution Depth Maps”. In: *Proceedings of the 13. IAPR International Conference on Machine Vision Applications, MVA 2013, Kyoto, Japan, May 20-23, 2013*. 2013, pp. 247–250.

[Mur94] R. M. Murray. *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[NE82] K. B. Nelson and J. H. Ellenberg. “Children Who ‘Outgrew’ Cerebral Palsy”. In: *Pediatrics* 69.5 (1982), pp. 529–536. ISSN: 0031-4005.

[Nel08] K. B. Nelson. “Causative factors in cerebral palsy”. In: *Clinical obstetrics and gynecology* 51.4 (2008), pp. 749–762.

[New+16] A. Newell et al. “Stacked Hourglass Networks for Human Pose Estimation”. In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe et al. Cham: Springer International Publishing, 2016, pp. 483–499. ISBN: 978-3-319-46484-8.

[Nis84] H. K. Nishihara. “Practical Real-Time Imaging Stereo Matcher”. In: *Optical Engineering* 23 (1984), pp. 23 –23 –10. DOI: 10.1117/12.7973334.

[Nov+17] I. Novak et al. “Early, accurate diagnosis and early intervention in cerebral palsy: Advances in diagnosis and treatment”. In: *JAMA Pediatrics* 171.9 (2017), pp. 897–907. DOI: 10.1001/jamapediatrics.2017.1689.

[NW06] J. Nocedal and S. J. Wright. *Numerical optimization*. 2006.

[Ohg+08] S. Ohgi et al. “Time Series Analysis of Spontaneous Upper-Extremity Movements of Premature Infants With Brain Injuries”. In: *Physical Therapy* 88.9 (2008), pp. 1022–1033. DOI: 10.2522/ptj.20070171.

[Ols+14] M. D. Olsen et al. “Body Part Tracking of Infants”. In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 2167–2172. DOI: 10.1109/ICPR.2014.377.

[Ols+15] M. D. Olsen et al. “Model-Based Motion Tracking of Infants”. In: *Computer Vision - ECCV 2014 Workshops*. Ed. by L. Agapito et al. Cham: Springer International Publishing, 2015, pp. 673–685. ISBN: 978-3-319-16199-0.

[Orl+18] S. Orlandi et al. “Detection of Atypical and Typical Infant Movements using Computer-based Video Analysis”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 3598–3601. DOI: 10.1109/EMBC.2018.8513078.

[Osk+13] M. Oskoui et al. “An update on the prevalence of cerebral palsy: a systematic review and meta-analysis”. In: *Developmental Medicine & Child Neurology* 55.6 (2013), pp. 509–519.

[Ozu+07] M. Ozuysal et al. “Fast Keypoint Recognition in Ten Lines of Code”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383123.

[Pal+12] R. J. Palisano et al. “Participation-based therapy for children with physical disabilities”. In: *Disability and Rehabilitation* 34.12 (2012). PMID: 22080765, pp. 1041–1052. DOI: 10.3109/09638288.2011.628740.

[Pal+97] R. Palisano et al. “Development and reliability of a system to classify gross motor function in children with cerebral palsy”. In: *Developmental Medicine & Child Neurology* 39.4 (1997), pp. 214–223. DOI: 10.1111/j.1469-8749.1997.tb07414.x.

[Pan08] N. Paneth. “Establishing the diagnosis of cerebral palsy”. In: *Clinical obstetrics and gynecology* 51.4 (2008), pp. 742–748.

[Pap+17] G. Papandreou et al. “Towards Accurate Multi-person Pose Estimation in the Wild”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3711–3719. DOI: 10.1109/CVPR.2017.395.

[Par+17] S. Park et al. “Accurate and Efficient 3D Human Pose Estimation Algorithm Using Single Depth Images for Pose Analysis in Golf”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, pp. 105–113. DOI: 10.1109/CVPRW.2017.19.

[Pav+17] G. Pavlakos et al. “Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 00. 2017, pp. 1263–1272. DOI: 10.1109/CVPR.2017.139.

[Pav+18] G. Pavlakos et al. “Ordinal Depth Supervision for 3D Human Pose Estimation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7307–7316. DOI: 10.1109/CVPR.2018.00763.

[Per+01] K. L. Perell et al. “Fall Risk Assessment MeasuresAn Analytic Review”. In: *The Journals of Gerontology: Series A* 56.12 (2001), pp. M761–M766. DOI: 10.1093/gerona/56.12.M761.

[Per+14] F. Perbet et al. “Human Body Shape Estimation Using a Multi-resolution Manifold Forest”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 668–675. DOI: 10.1109/CVPR.2014.91.

[Phi+14] H. Philippi et al. “Computer-based analysis of general movements reveals stereotypies predicting cerebral palsy”. In: *Developmental Medicine & Child Neurology* 56.10 (2014), pp. 960–967.

[Pis+16] L. Pishchulin et al. “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4929–4937. DOI: 10.1109/CVPR.2016.533.

[Pis+17] L. Pishchulin et al. “Building statistical shape spaces for 3D human modeling”. In: *Pattern Recognition* 67 (2017), pp. 276 –286. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.02.018.

[Pla+10] C. Plagemann et al. “Real-time identification and localization of body parts from depth images”. In: *2010 IEEE International Conference on Robotics and Automation*. 2010, pp. 3108–3113. DOI: 10.1109/ROBOT.2010.5509559.

[PM+13] G. Pons-Moll et al. “Metric Regression Forests for Human Pose Estimation”. In: *Proceedings of the British Machine Vision Conference*. BMVA Press, 2013.

[PM+15a] G. Pons-Moll et al. “Dyna: A Model of Dynamic Human Shape in Motion”. In: *ACM Trans. Graph.* 34.4 (July 2015), 120:1–120:14. ISSN: 0730-0301. DOI: 10.1145/2766993.

[PM+15b] G. Pons-Moll et al. “Metric Regression Forests for Correspondence Estimation”. In: *International Journal of Computer Vision* 113.3 (2015), pp. 163–175. ISSN: 1573-1405. DOI: 10.1007/s11263-015-0818-9.

[PM+17] G. Pons-Moll et al. “ClothCap: Seamless 4D Clothing Capture and Retargeting”. In: *ACM Trans. Graph.* 36.4 (July 2017), 73:1–73:15. ISSN: 0730-0301. DOI: 10.1145/3072959.3073711.

[PMR11] G. Pons-Moll and B. Rosenhahn. “Model-Based Pose Estimation”. In: *Visual Analysis of Humans: Looking at People*. Ed. by T. B. Moeslund et al. London: Springer London, 2011, pp. 139–170. ISBN: 978-0-85729-997-0. DOI: 10.1007/978-0-85729-997-0_9.

[PR91] D. Podsiadlo and S. Richardson. “The Timed “Up & Go”: A Test of Basic Functional Mobility for Frail Elderly Persons”. In: *Journal of the American Geriatrics Society* 39.2 (1991), pp. 142–148. DOI: 10.1111/j.1532-5415.1991.tb01616.x.

[Pre+97] H. F. Prechtl et al. “An early marker for neurological deficits after perinatal brain lesions”. In: *The Lancet* 349.9062 (1997), pp. 1361 – 1363. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(96)10182-3.

[Pre74] H. F. Prechtl. “The behavioural states of the newborn infant (a review)”. In: *Brain research* 76.2 (1974), pp. 185–212.

[Pre90] H. Prechtl. “Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction”. In: *Early Human Development* 23.3 (1990), pp. 151 –158. ISSN: 0378-3782. DOI: https://doi.org/10.1016/0378-3782(90)90011-7.

[Qam+18] A. Qammaz et al. “A Hybrid Method for 3D Pose Estimation of Personalized Human Body Models”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Vol. 00. 2018, pp. 456–465. DOI: 10.1109/WACV.2018.00056.

[Rah+12] H. Rahmati et al. “Kernel-Based Object Tracking for Cerebral Palsy Detection”. In: *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). 2012, p. 1.

[Rah+14a] H. Rahmati et al. “Video-based early cerebral palsy prediction using motion segmentation”. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 3779–3783. DOI: 10.1109/EMBC.2014.6944446.

[Rah+14b] H. Rahmati et al. “Motion Segmentation with Weak Labeling Priors”. In: *Pattern Recognition*. Ed. by X. Jiang et al. Cham: Springer International Publishing, 2014, pp. 159–171. ISBN: 978-3-319-11752-2.

[Rah+15a] H. Rahmati et al. “Frequency-based features for early cerebral palsy prediction”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 5187–5190.

[Rah+15b] H. Rahmati et al. “Weakly supervised motion segmentation with particle matching”. In: *Computer Vision and Image Understanding* 140 (2015), pp. 30–42.

[Rah+16] H. Rahmati et al. "Frequency Analysis and Feature Reduction Method for Prediction of Cerebral Palsy in Young Infants". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 24.11 (2016), pp. 1225–1234. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2016.2539390.

[Ram+12] V. Ramakrishna et al. "Reconstructing 3D Human Pose from 2D Image Landmarks". In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 573–586. ISBN: 978-3-642-33765-9.

[Rey+11] M. Reynolds et al. "Capturing Time-of-Flight data with confidence". In: *CVPR 2011*. 2011, pp. 945–952. DOI: 10.1109/CVPR.2011.5995550.

[Rho+18] H. Rhodin et al. "Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation". In: (2018). Ed. by V. Ferrari et al., pp. 765–782.

[Rob+02] K. M. Robinette et al. *Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report. Volume 1. Summary.* Tech. rep. SYTRONICS INC DAYTON OH, 2002.

[Rog+17] G. Rogez et al. "LCR-Net: Localization-Classification-Regression for Human Pose". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1216–1224. DOI: 10.1109/CVPR.2017.134.

[Rom+08] D. M. M. Romeo et al. "Early neurologic assessment in preterm-infants: Integration of traditional neurologic examination and observation of general movements". In: *European Journal of Paediatric Neurology* 12.3 (2008), pp. 183 –189. ISSN: 1090-3798. DOI: https://doi.org/10.1016/j.ejpn.2007.07.008.

[Rom+17] J. Romero et al. "Embodied Hands: Modeling and Capturing Hands and Bodies Together". In: *ACM Trans. Graph.* 36.6 (Nov. 2017), 245:1–245:17. ISSN: 0730-0301. DOI: 10.1145/3130800.3130883.

[Ros+07] P. Rosenbaum et al. "A report: the definition and classification of cerebral palsy April 2006". In: *Dev Med Child Neurol Suppl* 109.suppl 109 (2007), pp. 8–14.

[Rus+15] O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252. ISSN: 1573-1405. DOI: 10.1007/s11263-015-0816-y.

[Sah+15] S. Sahu et al. "Malnutrition among under-five children in India and strategies for control". In: *Journal of Natural Science, Biology and Medicine* 6.1 (2015), pp. 18–23. DOI: 10.4103/0976-9668.149072.

[SC+00] A. Shumway-Cook et al. "Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test". In: *Physical Therapy* 80.9 (2000), pp. 896–903. DOI: 10.1093/ptj/80.9.896.

[Sch+11] L. A. Schwarz et al. “Estimating human 3D pose from Time-of-Flight images based on geodesic distances and optical flow”. In: *Face and Gesture 2011*. 2011, pp. 700–706. DOI: 10.1109/FG.2011.5771333.

[Sci+17] G. Sciortino et al. “On the Estimation of Children’s Poses”. In: *Image Analysis and Processing - ICIAP 2017*. Ed. by S. Battiato et al. Cham: Springer International Publishing, 2017, pp. 410–421. ISBN: 978-3-319-68548-9.

[See+14] T. Seel et al. “IMU-Based Joint Angle Measurement for Gait Analysis”. In: *Sensors* 14.4 (2014), pp. 6891–6909. ISSN: 1424-8220. DOI: 10.3390/s140406891.

[Sel+16] E. Sellier et al. “Decreasing prevalence in cerebral palsy: a multi-site European population-based study, 1980 to 2003”. In: *Developmental Medicine & Child Neurology* 58.1 (2016), pp. 85–92.

[Ser+16] M. M. Serrano et al. “Lower limb pose estimation for monitoring the kicking patterns of infants”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016, pp. 2157–2160. DOI: 10.1109/EMBC.2016.7591156.

[Sha+14] A. Shapiro et al. “Rapid avatar capture and simulation using commodity depth sensors”. In: *Computer Animation and Virtual Worlds* 25.3-4 (2014), pp. 201–211.

[Sha08] T. Sharp. “Implementing Decision Trees and Forests on a GPU”. In: *Computer Vision – ECCV 2008*. Ed. by D. Forsyth et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 595–608. ISBN: 978-3-540-88693-8.

[Shi+17] S. S. Shivakumar et al. “Stereo 3D tracking of infants in natural play conditions”. In: *2017 International Conference on Rehabilitation Robotics (ICORR)*. 2017, pp. 841–846. DOI: 10.1109/ICORR.2017.8009353.

[Sho+11] J. Shotton et al. “Real-time human pose recognition in parts from single depth images”. In: *CVPR 2011*. 2011, pp. 1297–1304. DOI: 10.1109/CVPR.2011.5995316.

[Sig+09] L. Sigal et al. “HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion”. In: *International Journal of Computer Vision* 87.1 (2009), p. 4. ISSN: 1573-1405. DOI: 10.1007/s11263-009-0273-6.

[Sig14] L. Sigal. “Human pose estimation”. In: *Computer Vision*. Springer, 2014, pp. 362–370.

[Sim+17] T. Simon et al. “Hand Keypoint Detection in Single Images Using Multiview Bootstrapping”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4645–4653. DOI: 10.1109/CVPR.2017.494.

[Sin+16] S. Sinha et al. “Accurate upper body rehabilitation system using Kinect”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016, pp. 4605–4609. DOI: 10.1109/EMBC.2016.7591753.

[Sin+17a] V. Singh et al. “DARWIN: Deformable Patient Avatar Representation With Deep Image Network”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017*. Ed. by M. Descoteaux et al. Cham: Springer International Publishing, 2017, pp. 497–504. ISBN: 978-3-319-66185-8.

[Sin+17b] S. Sinha et al. “Accurate estimation of joint motion trajectories for rehabilitation using Kinect”. In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2017, pp. 3864–3867. DOI: 10.1109/EMBC.2017.8037700.

[SMT03] H. Seo and N. Magnenat-Thalmann. “An Automatic Modeling of Human Bodies from Sizing Parameters”. In: *Proceedings of the 2003 Symposium on Interactive 3D Graphics*. I3D ’03. Monterey, California: ACM, 2003, pp. 19–26. ISBN: 1-58113-645-5. DOI: 10.1145/641480.641487.

[SP10] M. Singh and D. J. Patterson. “Involuntary gesture recognition for predicting cerebral palsy in high-risk infants”. In: *Wearable Computers (ISWC), 2010 International Symposium on*. IEEE. 2010, pp. 1–8.

[Spi+15] A. Spittle et al. “Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants”. In: *Cochrane Database of Systematic Reviews* 11 (2015).

[Spi+16] A. Spittle et al. “The Baby Moves prospective cohort study protocol: using a smartphone application with the General Movements Assessment to predict neurodevelopmental outcomes at age 2 years for extremely preterm or extremely low birthweight infants”. In: *BMJ Open* 6.10 (2016). ISSN: 2044-6055. DOI: 10.1136/bmjopen-2016-013446.

[SS15] A. Schick and R. Stiefelhagen. “3D Pictorial Structures for Human Pose Estimation with Supervoxels”. In: *2015 IEEE Winter Conference on Applications of Computer Vision*. 2015, pp. 140–147. DOI: 10.1109/WACV.2015.26.

[Sta+00] F. J. Stanley et al. *Cerebral palsies: epidemiology and causal pathways*. 151. Cambridge University Press, 2000.

[Sta+12] A. Stahl et al. “An Optical Flow-Based Method to Predict Infantile Cerebral Palsy”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20.4 (2012), pp. 605–614. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2012.2195030.

[Sto+11] C. Stoll et al. “Fast articulated motion tracking using a sums of Gaussians body model”. In: *2011 International Conference on Computer Vision*. 2011, pp. 951–958. DOI: 10.1109/ICCV.2011.6126338.

[Str+12a] M. Straka et al. “Rapid Skin: Estimating the 3D human pose and shape in real-time”. In: *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*. IEEE. 2012, pp. 41–48.

[Str+12b] M. Straka et al. “Simultaneous Shape and Pose Adaption of Articulated Models Using Linear Optimization”. In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 724–737. ISBN: 978-3-642-33718-5.

[Stø+17] R. Støen et al. “Computer-based video analysis identifies infants with absence of fidgety movements”. In: *Pediatric research* 82.4 (2017), pp. 665–670. ISSN: 0031-3998. DOI: 10.1038/pr.2017.121.

[Sun+12] M. Sun et al. “Conditional regression forests for human pose estimation”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 3394–3401. DOI: 10.1109/CVPR.2012.6248079.

[Sun+14] B. Sun et al. “Human gait modeling and gait analysis based on Kinect”. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 3173–3178. DOI: 10.1109/ICRA.2014.6907315.

[Sun+17] X. Sun et al. “Compositional Human Pose Regression”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2621–2630. DOI: 10.1109/ICCV.2017.284.

[Tac+17] U. Tacke et al. “Entwicklungsneurologie – vernetzte Medizin und neue Perspektiven”. In: *Der Nervenarzt* 88 (2017), pp. 1395–1401. ISSN: 1433-0407. DOI: 10.1007/s00115-017-0436-6.

[Tao+15] L. Tao et al. “A comparative study of pose representation and dynamics modelling for online motion quality assessment”. In: *Computer Vision and Image Understanding* 11 (2015).

[Tay+12] J. Taylor et al. “The Vitruvian Manifold: Inferring dense correspondences for one-shot human pose estimation”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 103–110. DOI: 10.1109/CVPR.2012.6247664.

[TB14] M. A. Timo Breuer Christoph Bodensteiner. “Low-cost commodity depth sensor comparison and accuracy analysis”. In: vol. 9250. 2014, pp. 9250 –9250 –10. DOI: 10.1117/12.2067155.

[Tek+16] B. Tekin et al. “Direct Prediction of 3D Body Poses from Motion Compensated Sequences”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 991–1000. DOI: 10.1109/CVPR.2016.113.

[Tek+17] B. Tekin et al. “Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 3961–3970. DOI: 10.1109/ICCV.2017.425.

[Tom+14] J. J. Tompson et al. “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 1799–1807.

[Tom+17] D. Tome et al. “Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5689–5698. DOI: 10.1109/CVPR.2017.603.

[Ton+12] J. Tong et al. “Scanning 3D Full Human Bodies Using Kinects”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.4 (2012), pp. 643–650. ISSN: 1077-2626. DOI: 10.1109/TVCG.2012.56.

[TS14] A. Toshev and C. Szegedy. “DeepPose: Human Pose Estimation via Deep Neural Networks”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1653–1660. DOI: 10.1109/CVPR.2014.214.

[Tso+14] A. Tsoli et al. “Breathing Life into Shape: Capturing, Modeling and Animating 3D Human Breathing”. In: *ACM Trans. Graph.* 33.4 (July 2014), 52:1–52:11. ISSN: 0730-0301. DOI: 10.1145/2601097.2601225.

[Tun+17] H.-Y. Tung et al. “Self-supervised Learning of Motion Capture”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5236–5246.

[Var+18] G. Varol et al. “BodyNet: Volumetric Inference of 3D Human Body Shapes”. In: *Computer Vision – ECCV 2018*. Ed. by V. Ferrari et al. Cham: Springer International Publishing, 2018, pp. 20–38. ISBN: 978-3-030-01234-2.

[Wei+11] A. Weiss et al. “Home 3D body scans from noisy image and range data”. In: *2011 International Conference on Computer Vision*. 2011, pp. 1951–1958. DOI: 10.1109/ICCV.2011.6126465.

[Wei+16] S. Wei et al. “Convolutional Pose Machines”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 00. 2016, pp. 4724–4732. DOI: 10.1109/CVPR.2016.511.

[Wel+17] T. Welschehold et al. “Learning mobile manipulation actions from human demonstrations”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017, pp. 3196–3201. DOI: 10.1109/IROS.2017.8206152.

[Whi96] M. W. Whittle. “Clinical gait analysis: A review”. In: *Human Movement Science* 15.3 (1996), pp. 369 –387. ISSN: 0167-9457. DOI: https://doi.org/10.1016/0167-9457(96)00006-1.

[Wit+99] I. H. Witten et al. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, May 1999.

[WS06] J. Winn and J. Shotton. “The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects”. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 1. 2006, pp. 37–44. DOI: 10.1109/CVPR.2006.305.

[Wuh+14] S. Wuhrer et al. “Estimation of human body shape and posture under clothing”. In: *Computer Vision and Image Understanding* 127 (2014), pp. 31–42.

[Xia+18] S. Xia et al. “Cascaded 3D Full-Body Pose Regression from Single Depth Image at 100 FPS”. In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE. 2018, pp. 431–438.

[Ye+12] G. Ye et al. “Performance Capture of Interacting Characters with Handheld Kinects”. In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 828–841. ISBN: 978-3-642-33709-3.

[Ye+13] M. Ye et al. “A Survey on Human Motion Analysis from Depth Data”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications: Dagstuhl 2012 Seminar on Time-of-Flight Imaging and GCPR 2013 Workshop on Imaging New Modalities*. Ed. by M. Grzegorzek et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 149–187. ISBN: 978-3-642-44964-2. DOI: 10.1007/978-3-642-44964-2_8.

[Yu+17] T. Yu et al. “BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 910–919. DOI: 10.1109/ICCV.2017.104.

[Yu+18] T. Yu et al. “DoubleFusion: Real-Time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7287–7296. DOI: 10.1109/CVPR.2018.00761.

[YY14] M. Ye and R. Yang. “Real-Time Simultaneous Pose and Shape Estimation for Articulated Objects Using a Single Depth Camera”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2353–2360. DOI: 10.1109/CVPR.2014.301.

[ZB15] S. Zuffi and M. J. Black. "The Stitched Puppet: A graphical model of 3D human shape and pose". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3537–3546. DOI: 10.1109/CVPR.2015.7298976.

[Zha+14] Q. Zhang et al. "Quality Dynamic Human Body Modeling Using a Single Low-Cost Depth Camera". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 676–683. DOI: 10.1109/CVPR.2014.92.

[Zha+17] C. Zhang et al. "Detailed, Accurate, Human Shape Estimation from Clothed 3D Scan Sequences". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5484–5493. DOI: 10.1109/CVPR.2017.582.

[Zhe+14] J. Zheng et al. "SCAPE-based human performance reconstruction". In: *Computers & Graphics* 38 (2014), pp. 191 –198. ISSN: 0097-8493. DOI: https://doi.org/10.1016/j.cag.2013.10.023.

[Zho+16] X. Zhou et al. "Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4966–4975. DOI: 10.1109/CVPR.2016.537.

[Zim+18] C. Zimmermann et al. "3D Human Pose Estimation in RGBD Images for Robotic Task Learning". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018, pp. 1986–1992. DOI: 10.1109/ICRA.2018.8462833.

[Zuf+17] S. Zuffi et al. "3D Menagerie: Modeling the 3D Shape and Pose of Animals". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5524–5532. DOI: 10.1109/CVPR.2017.586.

[Zuf+18] S. Zuffi et al. "Lions and Tigers and Bears: Capturing Non-rigid, 3D, Articulated Shape from Images". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3955–3963. DOI: 10.1109/CVPR.2018.00416.

[Kho11] K. Khoshelham. "Accuracy Analysis of Kinect Depth Data". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 3812 (Sept. 2011), pp. 133–138. DOI: 10.5194/isprsarchives-XXXVIII-5-W12-133-2011.

[Ope] OpenPose library. Accessed June 2018.

[Sch+15] F. Schroff et al. "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[Sze+15] C. Szegedy et al. “Going deeper with convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

A. Overview of data sets and body models

We present tables containing the data sets we use for evaluation in Tab. A.1, and the different body models in Tab. A.2.

Table A.1.: Evaluation data sets used in different chapters.

Data set	Chapter	Approach	# seq	# frames	Ground truth	Info
PDT	Sec. 4.2.3	Ferns	20	27K	Marker-based MoCap	Adults, Kinect V1
FernSeq1	Sec. 4.2.3	Ferns	1	1082	Manual annotation	Kinect V2
MINI-RGBD	Sec. 4.2.3, Sec. 4.3	Ferns	12	12K	SMIL	Synthetic
FernMultiView	Sec. 4.3	Ferns extension	1	72	Manual annotation	Kinect V2
FernSeq2	Sec. 4.3	Ferns extension	3	5500	$SMPL_B$ (prel.)	Kinect V1 + V2
SMILTrain	Sec. 5.4	SMIL	37	200K	Manual inspection	Kinect V1

Table A.2.: Models and usages.

Model	Usage	Chapters
makeHuman	Training data generation	Sec. 4.2.2, Sec. 4.3.2
SMPL	Adult model, foundation of SMIL	Sec. 5.1.2
$SMPL_B$ (prel.)	Evaluation: Ground truth generation	Sec. 4.3
SMIL	Motion capture + evaluation data generation	Sec. 5.4

B. MINI-RGBD sequences

Figure B.1.: Sequence 1, level: easy, samples from every 50 frames.

Figure B.2.: Sequence 2, level: easy, samples from every 50 frames.

Figure B.3.: Sequence 3, level: easy, samples from every 50 frames.

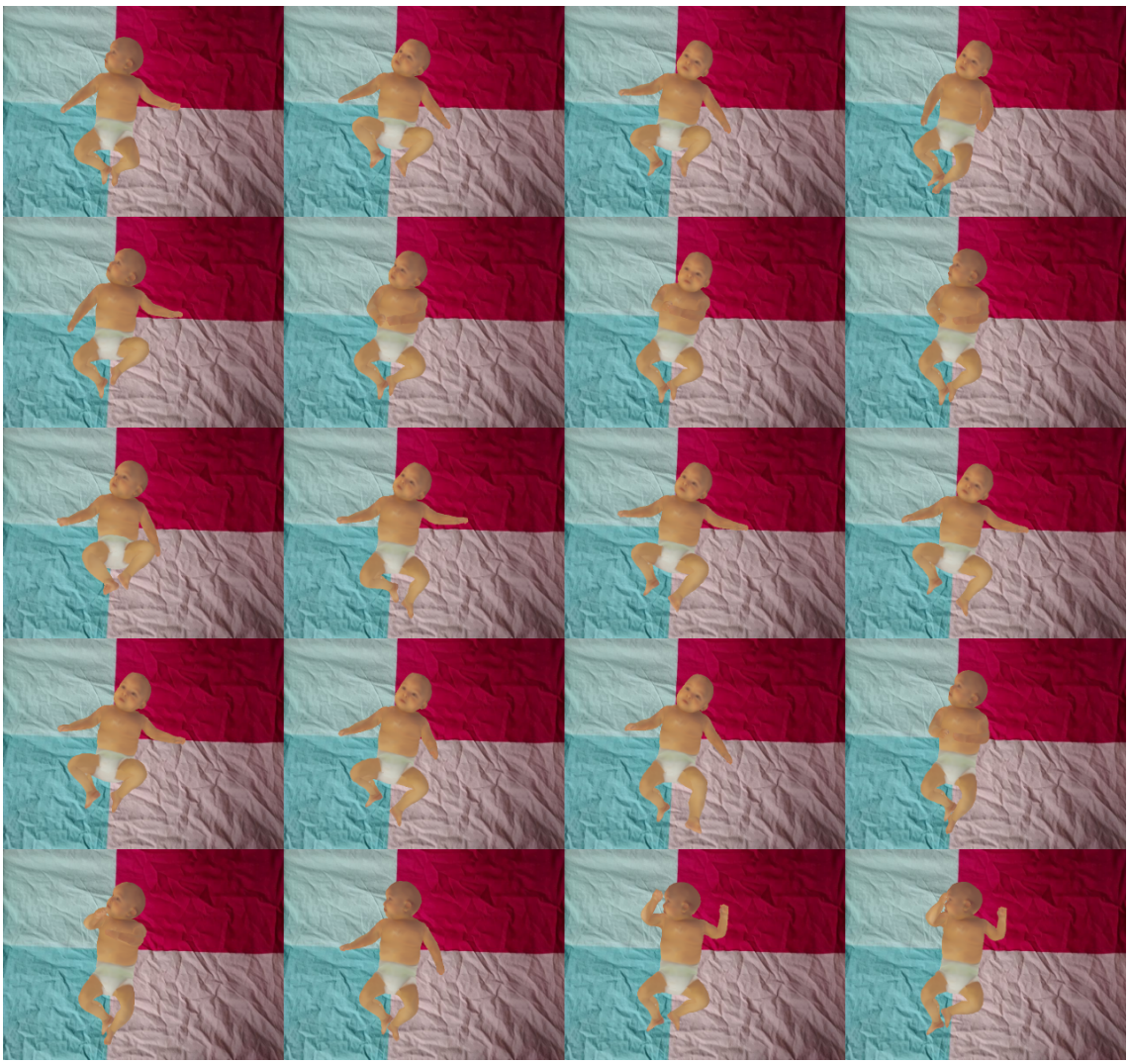


Figure B.4.: Sequence 4, level: easy, samples from every 50 frames.

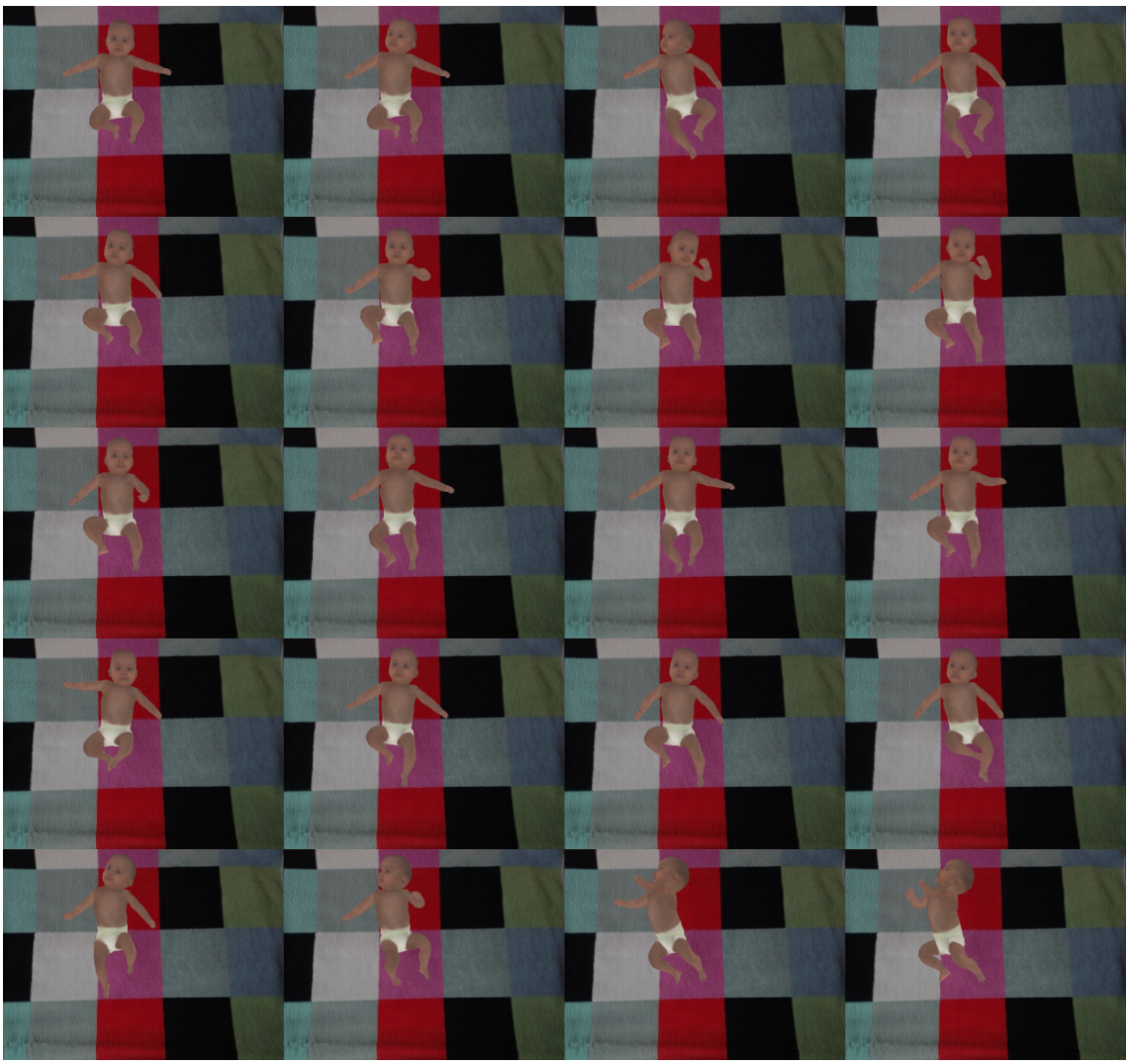


Figure B.5.: Sequence 5, level: medium, samples from every 50 frames.

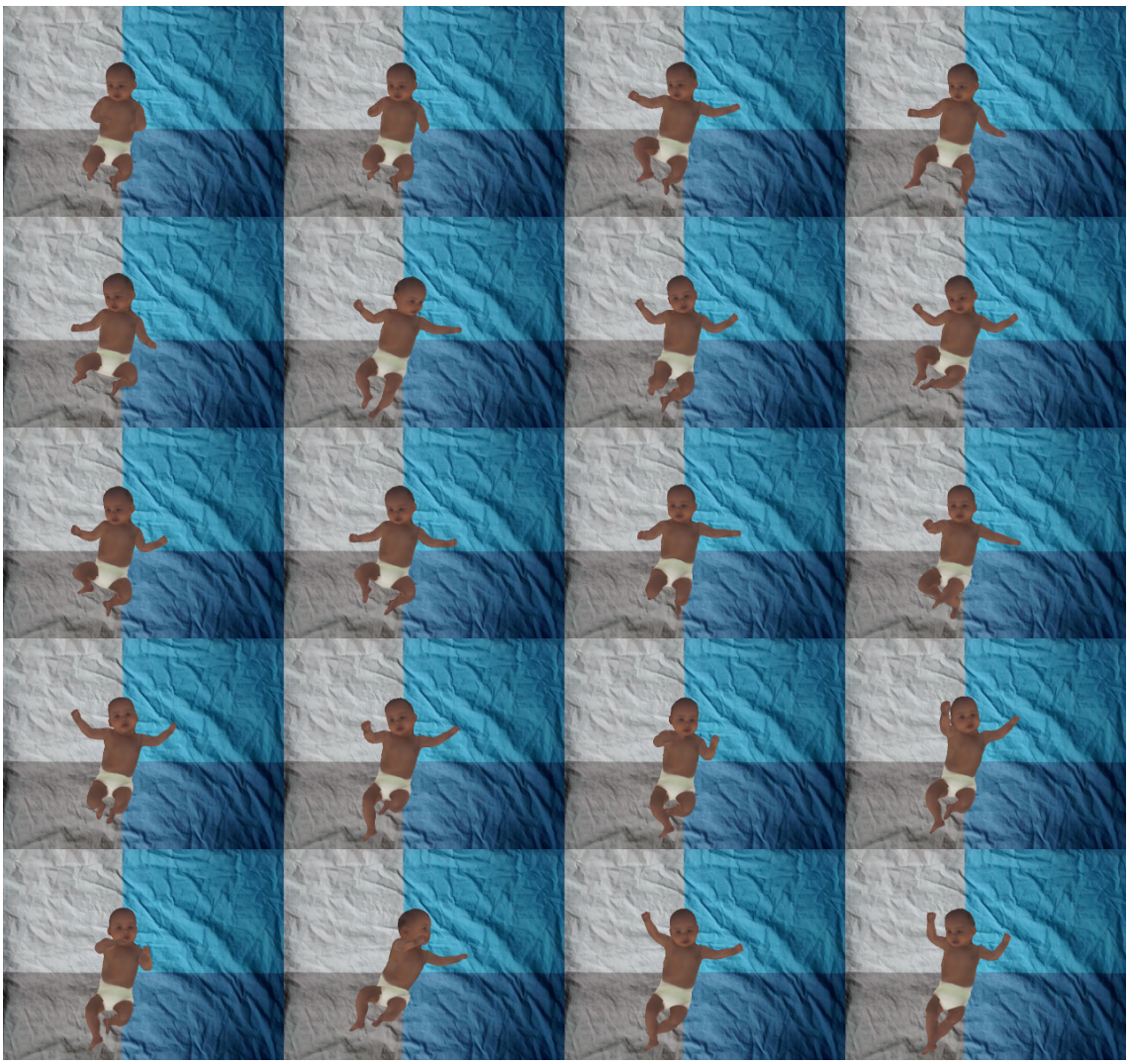


Figure B.6.: Sequence 6, level: medium, samples from every 50 frames.

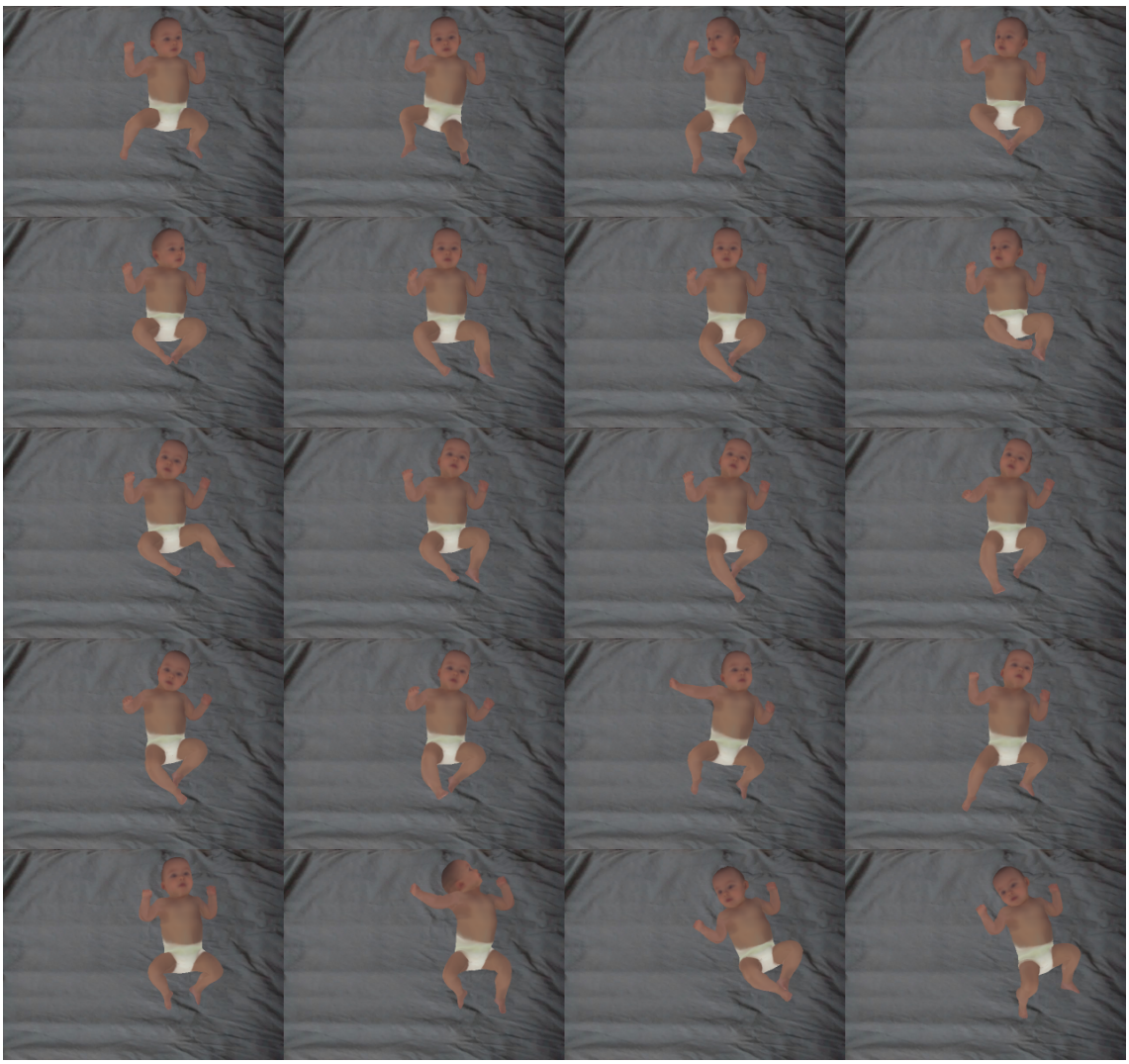


Figure B.7.: Sequence 7, level: medium, samples from every 50 frames.

Figure B.8.: Sequence 8, level: medium, samples from every 50 frames.

Figure B.9.: Sequence 9, level: medium, samples from every 50 frames.

Figure B.10.: Sequence 10, level: difficult, samples from every 50 frames.

Figure B.11.: Sequence 11, level: difficult, samples from every 50 frames.

Figure B.12.: Sequence 12, level: difficult, samples from every 50 frames.

C. Motion word samples

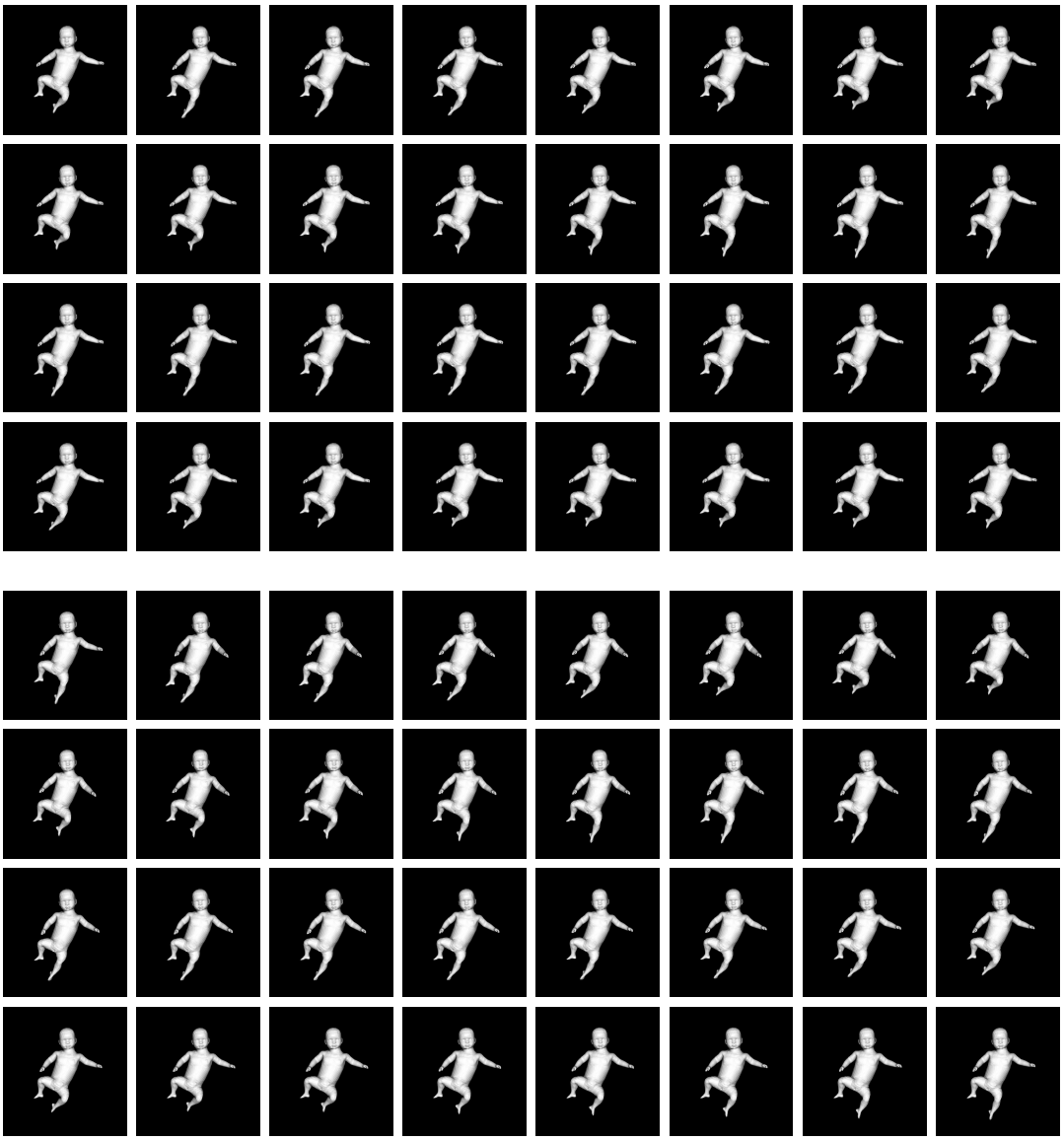


Figure C.1.: Top: Motion word sample with duration of 32 frames. Bottom: Best matching motion word, i.e., word with lowest dynamic time warping distance to input sample, from the same sequence.