Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set

Nikolas Hesse[1*][0000−0003−1141−0614], Christoph Bodensteiner[1], Michael Arens[1][0000−0002−7857−0332], Ulrich G. Hofmann[2][0000−0002−6264−3701], Raphael Weinberger[3], and A. Sebastian Schroeder[3]

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Ettlingen, Germany
[2] University Medical Center Freiburg, Faculty of Medicine, University of Freiburg, Germany
[3] Ludwig Maximilian University, Hauner Children's Hospital, Munich, Germany

Abstract. Assessment of spontaneous movements of infants lets trained experts predict neurodevelopmental disorders like cerebral palsy at a very young age, allowing early intervention for affected infants. An automated motion analysis system requires to accurately capture body movements, ideally without markers or attached sensors to not affect the movements of infants. A vast majority of recent approaches for human pose estimation focuses on adults, leading to a degradation of accuracy if applied to infants. Hence, multiple systems for *infant* pose estimation have been developed. Due to the lack of publicly available benchmark data sets, a standardized evaluation, let alone a comparison of different approaches is impossible. We fill this gap by releasing the Moving INfants In RGB-D (MINI-RGBD)[†] data set, created using the recently introduced *Skinned Multi-Infant Linear* body model (SMIL). We map real infant movements to the SMIL model with realistic shapes and textures, and generate RGB and depth images with precise ground truth 2D and 3D joint positions. We evaluate our data set with state-of-the-art methods for 2D pose estimation in RGB images and for 3D pose estimation in depth images. Evaluation of 2D pose estimation results in a PCKh rate of 88.1% and 94.5% (depending on correctness threshold), and PCKh rates of 64.2%, respectively 90.4% for 3D pose estimation. We hope to foster research in medical infant motion analysis to get closer to an automated system for early detection of neurodevelopmental disorders.

Keywords: Motion analysis, infants, pose estimation, RGB-D, data set, cerebral palsy

1 Introduction

Advances in computer vision and the widespread availability of low-cost RGB-D sensors have paved the way for novel applications in medicine [28], e.g. Alzheimer's

*nikolas.hesse@iosb.fraunhofer.de
†Data set available for research purposes at http://s.fhg.de/mini-rgbd

disease assessment [17], quantification of multiple sclerosis progression [22], as well as gait [43] or motion analysis [8]. For the latter, accurately capturing human movements is the fundamental step. Human pose estimation from RGB or depth images, especially using convolutional neural networks (CNNs), currently receives a lot of attention from the research community [5,44,12,45]. However, research is largely focused on adults. The ability to accurately capture *infant* movements is fundamental for an automated assessment of motor development.

In his pioneering work, Prechtl found that the quality of spontaneous movements of infants is a good marker for detecting impairments of the young nervous system [33]. This discovery led to the development of the General Movements Assessment (GMA) [33,32,11] which allows detection of neurodevelopmental disorders at a very young age. An automated system, relying on data captured by cheap RGB or RGB-D cameras could enable the widespread screening of all infants at risk of motion disorders. This would allow early intervention for affected children, which is assumed to improve outcome [41].

The application of state-of-the-art adult pose estimation systems to *children* was recently studied [36]. Authors found that children are underrepresented in most widely used benchmark data sets. Their analysis revealed that pose estimation accuracy decreases when approaches that were trained on adult data are applied to children. To mitigate this problem, they create a data set that was collected from internet videos, comprising 1176 images of 104 different children in unconstrained settings. For the evaluation of pose estimation approaches, they manually annotate 2D joint positions of 22 body keypoints.

Why does no *infant* data set exist? As mentioned above, computer vision research is primarily focused on adults. Reasons might include the higher number of potential applications for adults, with infant motion analysis being more of a niche application. Furthermore, it is not easy to generate reliable ground truth for infants. Manual annotation, especially in 3D, is error prone and cumbersome. Capturing ground truth with standard motion capture systems using markers may affect the infants' behavior, while suffering from problems with occlusions [27]. Researchers have used robotics to generate ground truth data [37,25], but reproducing the complexity of real infant movements is not possible with justifiable efforts. Other than that, laws are more strict concerning the privacy of children, since infants can not decide whether or not they want their image to be published. This makes the creation of a data set containing real infant images more challenging.

We fill this gap by creating an RGB-D data set for the evaluation of motion capture systems for infant motion analysis using the recently introduced *Skinned Multi-Infant Linear* model (SMIL) [14]. Our *Moving INfants In RGB-D* (MINI-RGBD) data set contains realistic motion, realistic appearance, realistic shapes, and precise 2D and 3D ground truth, and covers a wide range of challenging motions. To preserve the privacy of infants, we generate new textures and shapes by averaging multiple textures and shapes of real infants. These are still highly realistic, yet do not show any existing infant. We map real infant movements to these new "synthetic infants" and render RGB and depth images to simulate

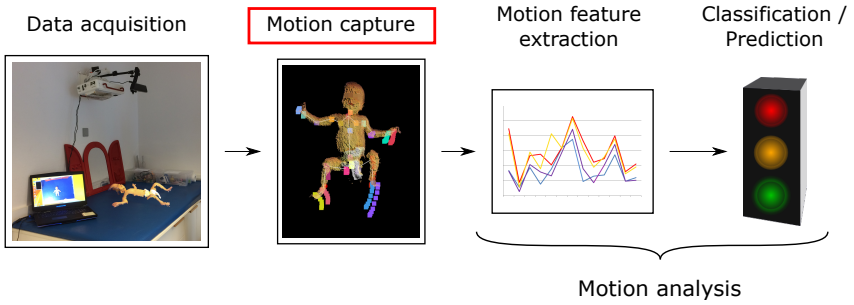


Fig. 1. Standard motion analysis pipeline. After data is acquired, motions are captured. Motion features are extracted and used to classify or predict the medical outcome. In this work, we focus on motion capture from RGB or depth images.

standard commodity RGB-D sensors. Our data set (described in Sec. 3), differs from the data set of [36] in multiple ways: (i) it contains infants up to the age of 7 months, (ii) we consider constrained settings for medical motion analysis, i.e. infants lie in supine position, facing the camera, (iii) we provide sequences of continuous motions instead of single frames, (v) we render data from a realistic 3D infant body model instead of annotating real images, (iv) we generate RGB and depth images, and (vi) we provide accurate 2D and 3D ground truth joint positions that are directly regressed from the model.

In the following, we review the state of the art in infant motion analysis and analyze evaluation procedures (Sec. 2). We describe the creation of our data set in Sec. 3, and present pose estimation baseline evaluations for RGB and RGB-D data in Sec. 4.

2 Medical Infant Motion Analysis - State of the Art

We review systems aiming at the automated prediction of cerebral palsy (CP) 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. 1). Motion features are extracted from captured movements, and used for training a classifier to predict the outcome. The reviewed systems report high sensitivity and specificity for CP prediction, mostly on study populations containing a small number of infants with confirmed CP diagnosis. Yet, the majority of approaches shows a considerable lack of evaluation of motion capture methods. The majority of approaches only scarcely evaluates the accuracy of 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. Of course, the non-existence of a public benchmark data set makes it hard to conduct an extensive evaluation of motion capture accuracy.

In this section, we present an overview of methods used for infant motion capture and how these are evaluated. The reader is referred to [26] for an extensive overview of the motion analysis stage of different approaches.

2.1 Wearable Motion Sensors

Although a recent study shows that wearable sensors do not seem to affect the leg movement frequency [18], they supposedly have a negative influence on the infant's content. 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) [20]. 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. Proposed systems using wearable sensors use wired [13] and wireless accelerometers [40,10,9], electromagnetical sensors [21,20], or a pressure mattress in combination with Inertial Measurement Units (IMU) [35]. In the following, we focus on vision-based approaches, and refer the reader to [7] for an overview of monitoring infant movements using wearable sensors.

2.2 Computer Vision for Capturing Movements

Cameras, opposed to motion sensors, are cheap, easy to use, require no setup or calibration, 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 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.

Video-based approaches. 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. They 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.

One of the first approaches towards automated CP prediction was introduced in 2006 by Meinecke et al. [27]. A commercial Vicon system uses 7 infrared cameras to track 20 markers, distributed across the infant's body. After an initial calibration procedure, the known marker positions on a biomechanical model are used to calculate the rotation of head and trunk, as well as the 3D positions of upper arms, forearms, thighs, lower legs, and feet from the tracked markers on the infant. The system is highly accurate, authors report errors 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, especially the markers of the upper extremities were occluded and therefore invisible to the cameras half of the measurement time. Attaching additional markers exceeded the system's capabilities and therefore, authors refrained from using motion features of upper extremities for CP prediction. 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.

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 [19]. They normalize the marker displacement data using the body size of the infant and smooth the resulting 2D position data. The accuracy of the capture system is not reported.

Machireddy et al. [25] present a hybrid system that 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 hands, legs, 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. The IMUs are synchronized with the camera, and the output of all sensors is fused using an extended Kalman filter. Ground truth for evaluation is generated by rotating a plywood model of a human arm using a drill, equipped with one marker and one IMU. Authors present plots of ground truth positions and estimated positions for a circular and a spiral motion. Exact numbers on accuracy are not presented.

Adde et al. take a more holistic approach [1]. Instead of tracking 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. Furthermore, 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. The accuracy of the system is not evaluated.

Stahl et al. use a motion tracking method based on optical flow between consecutive RGB frames [42]. They initialize points on a regular grid, distributed across the image, and track them over time. They evaluate the approach by manually selecting five points to be tracked from the grid as head, hands, and feet, and manually correct tracking errors. They display the result of their evaluation in one plot over 160 frames. Numbers on average accuracy are not given.

Rahmati et al. present an approach for motion segmentation using weak supervision [34]. Initialized by manual labeling, they track the body segmentation trajectories using optical flow fields. In case a trajectory ends due to fast motion or occlusion, they apply a particle matching algorithm for connecting a newly started trajectory for the same body segment. 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 segmenta-

Table 1. Summary of motion capture methods and corresponding evaluation of depth-based approaches for medical infant motion analysis. SD denotes standard deviation.

First author, year, reference	Method // tracked limbs	Ground truth (GT) generation // Reported avg. accuracy
Olsen 2014 [30]	Geodesic distances // 11 3D joint positions	Manual annotation (number of frames not specified) // 9 cm (extracted from plot)
Olsen 2014 [29]	Model-based tracking // 11 3D joint positions	Manual annotation (number of frames not specified) // 5 cm (SD: 3 cm) (extracted from plot)
Hesse 2015 [16]	Random ferns body part classifier // 21 3D joint positions	Manual annotation of 1082 frames // 4.1 cm
Hesse 2017 [15]	Random ferns (extension of [16]) // 21 3D joint positions	3D model fitting (+ visual verification), 5500 frames (3 seq.) // 1.2 cm (SD: 0.9 cm)
Hesse 2018 [14]	Model-based tracking // full body pose and shape	No GT, evaluation on 37 seq. (200K frames, $\sim$2 hours), pose errors determined from visual examination // average scan to model distance 2.51 mm, 34 pose errors lasting 90 s ($\approx$ 1.2%)
Serrano 2016 [37]	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 2017 [6]	Movement blobs // arms and legs	No evaluation
Shivakumar 2017 [38]	Optical flow + color-based segmentation // 3D positions of head, torso, hands, feet	Manual annotation of 60 frames // 8.21 cm, SD: 8.75 cm

tion. 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%.

Depth-based approaches With the introduction of low-cost RGB-D sensors, motion analysis approaches started taking advantage of depth information. The most well-known RGB-D camera is probably 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, 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. An overview of examined approaches is given in Tab. 1.

Olsen et al. transfer an existing pose estimation approach to infants [30]. The underlying assumption is that extremities have maximum geodesic distance to the body center. The body center is localized by filtering the infant's clothing color, based on a fixed threshold. They locate five anatomical extremities by finding points on the body farthest from the body center. Assuming a known body orientation, each of these points is assigned to one of the classes head, left / right hand, left / right foot, based on the spatial location and the orientation of the path to 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, they annotate 3D joint positions on an unspecified number of frames. 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 use a model-based approach for tracking eleven 3D joint positions [29]. They construct a human body model from simplistic shapes (cylinders, sphere, ellipsoid). After determining size parameters of the body parts, their previous method [30] is used for finding an initial pose. They fit the body model to the segmented infant point clouds that are computed from depth images. They optimize an objective function, defined by the difference of closest points from point cloud and model, with respect to the model pose parameters using the Levenberg-Marquardt algorithm. As in previous work, they evaluate the accuracy of their system on manually annotated 3D joint positions of an unspecified number of frames. The results are compared to their previous approach. Numbers are extracted from presented plots. The model-based system achieves an average joint position error of 5 cm (standard deviation (SD) 3 cm). Largest errors occur for right hand (7 cm) and stomach (6 cm).

Inspired by the approach used in the Kinect SDK, Hesse et al. propose a system for the estimation of 21 joint positions using a random ferns body part classifier [16]. A synthetic infant body model is used for rendering a large number of labeled depth images, from which a pixel-wise body part classifier based on binary depth comparison features is trained. 3D joint positions are calculated as the mean of all pixels belonging to each estimated body part. The system is trained on synthetic adult data and evaluated on the PDT benchmark data set containing adults. An average joint positions error of 13 cm is reported, compared to 9.6 cm for the Kinect SDK. The authors manually annotated 3D joint positions of an infant sequence consisting of 1082 frames. They report an average joint position error of 4.1 cm, with left hand (14.9 cm) and left shoulder (7.3 cm) showing the largest errors. They explain the errors with wrongly classified body parts for poses that were not included in the training set.

The approach is extended in [15] by including a feature selection step, generating more infant-like poses for training data, integrating kinematic chain constraints, and by applying PCA on torso pixels to correct for body rotations.

Ground truth joint positions are generated for 5500 frames of 3 sequences by fitting a body model and visually verifying the accuracy of the results. The best average error of the proposed methods is reported as 1.2 cm (SD 0.9 cm), compared to 1.8 cm (SD 3.1 cm) of the initial approach [16]. Additionally, a more strict evaluation metric, the worst-case accuracy, is applied. It denotes the percentage of frames for which all joint errors are smaller than a given threshold. For a threshold of 5 cm, 90% of frames are correct for [15], and 55% for [16], a threshold of 3cm decreases the accuracy to 50%, and less than 30%, respectively.

In recent work, Hesse et al. propose a model-based approach for estimating pose and shape of infants [14]. They learn an infant body model from RGB-D recordings and present an algorithm for fitting the model to the data. They optimize an objective function, consisting of scan to model distance, similar to [29], but add more terms, e.g. integrating prior probabilities of plausible shapes and poses. The average distance of scan points to model surface for 200K frames of 37 infant sequences (roughly 2 hours) is 2.51 mm (SD 0.21 mm). From manual inspection of all sequences, they report 18 failure cases in 7 sequences and 16 unnatural foot rotations lasting altogether 90 seconds, which corresponds to 1.2% of overall duration.

Serrano et al. track lower extremities using a leg model [37]. The approach is semi-automatic and requires some 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.

In [6], Cenci et al. use the difference image between two frames with a defined delay in between. After noise filtering, the difference image is segmented into motion blobs using a threshold. K-means clustering assigns each of the movement blobs to one of four different body parts (arms and legs). A state vector is generated for each frame, which contains information on which limb moves / does not move in this frame. There is no evaluation of the correctness of assigning blobs to limb classes.

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 [38]. After initially locating the body center based on a color threshold, an ellipse is fitted to the colored region and tracked. 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.

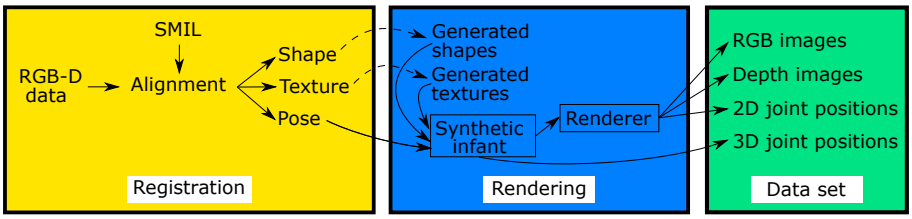


Fig. 2. Overview of data set creation pipeline. SMIL body model [14] is aligned to real infant RGB-D data. Subsets of shapes and textures are used for generating realistic, privacy preserving infant bodies. We animate the new "synthetic infants" with real movements (poses) from the registration stage. We render RGB and depth images, and create ground truth 2D and 3D joint positions to complete our new data set.

To summarize the evaluation protocols of reviewed approaches, comparison to previous work was limited to works of the same authors. Ground truth was mostly, if at all, generated by manual annotation of a small number of frames or by relying on robotics. This emphasizes the need for an infant benchmark RGB-D data set.

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

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, registration and rendering (see Fig. 2). Two samples of rendered images and joint positions are displayed in Fig. 3.

3.1 Registration

The capturing of shape and pose is achieved by registering SMIL to 12 RGB-D sequences of moving infants that were recorded in a children's hospital. Written informed consent was given by parents and ethics approval was obtained from Ludwig Maximilian University Munich. SMIL is based on SMPL [23], and shares the same properties. The model can be regarded as a parametric mapping, with pose and shape parameters serving as input, and output being a triangulated mesh of the shaped and posed infant, consisting of 6890 vertices. The model contains 23 body joints, each of which has three degrees of freedom (DOF). Together with 3 DOF for the global orientation this gives 72 pose parameters.

We follow the protocol of [14] to register the SMIL model to point clouds created from RGB-D sequences (which we will also call "scans"). We briefly recap the method and refer the reader to [14] for additional details.

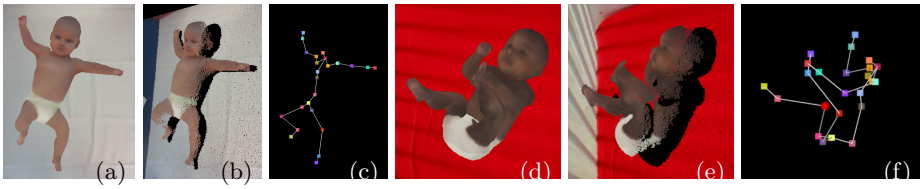


Fig. 3. Two samples from MINI-RGBD data set. (a) and (d): RGB image. (b) and (e): point cloud created from depth image. (c) and (f): ground truth skeleton. Viewpoint for (b), (c), (e), and (f) is slightly rotated to side. Best viewed in color.

To register the model to a scan, an objective function is optimized w.r.t. pose and shape parameters of SMIL. The function consists of the following terms: (i) distance between scan points and model mesh surface, (ii) a landmark error terms that penalizes distances between model landmarks projected to 2D and 2D body, face, and hand landmark estimates from RGB images using OpenPose library [5,39,44,31], (iii) a temporal pose smoothness term, (iv) a penalty for self intersections, (v) a term for keeping the back-facing model vertices close to, but not inside the background table, and (vi) prior probabilities on plausible shapes and poses. This results in a posed and shaped model mesh that describes the input point cloud data. The initialization frame is automatically selected based on 2D pose estimates. For the rest of the sequence, the resulting parameters of the last processed frame are used as initialization for the subsequent frame.

Going beyond the methods of [14], we generate one texture for each sequence, similar to [3,4]. 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 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. 4 (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 create a variety of realistic body shapes by averaging different subsets of shapes from the registration stage (Fig. 4 (b)).

3.2 Rendering

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

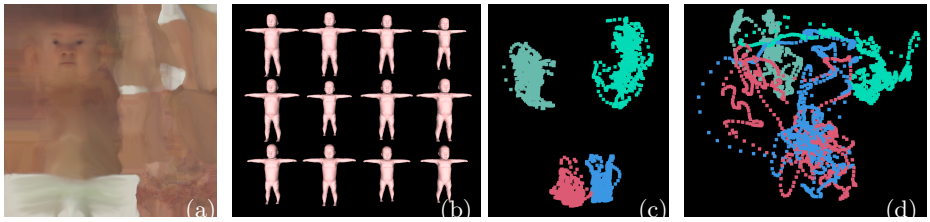


Fig. 4. (a) Sample of generated texture. (b) Generated shapes in T-pose. (c) Plotted joint positions from an "easy" sequence. Hand positions shown in light and dark green. Foot positions in red and blue. (d) Hand and foot positions for a "difficult" sequence. Color coding as in (c). Best viewed in color.

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 [24] 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 to Microsoft Kinect V1, which is the most frequently used sensor in approaches in Sec. 2, 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. 3 (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.

3.3 MINI-RGBD 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 half year of life, and we divide the sequences into different levels of difficulty (see Fig. 4 (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 skeletons. To properly compare these approaches, we add one frame in T-pose (extended arms and legs, cf. Fig. 4 (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.

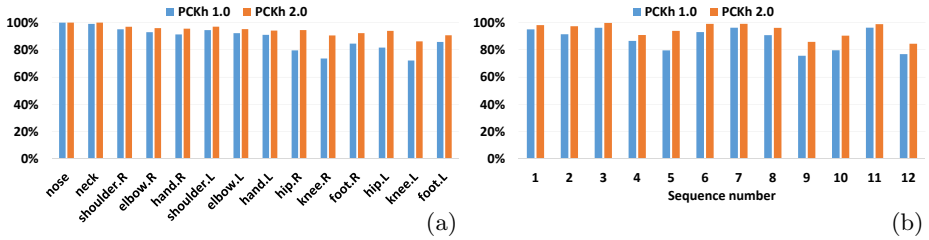


Fig. 5. RGB evaluation. Results for 2D pose estimation from OpenPose library. (a) Percentage of correct keypoints in relation to head segment length (PCKh) per joint. *PCKh 1.0* denotes a correctness threshold of one time head segment length, *PCKh 2.0* of twice the head segment length. (b) PCKh per sequence.

4 Evaluation

We provide baseline evaluations using state-of-the-art approaches for the RGB, as well as the RGB-D part of our MINI-RGBD data set.

4.1 2D Pose Estimation in RGB Images

We use a state-of-the-art adult RGB pose estimation system from OpenPose library [5,31] 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. 3.3), and add these offsets to the estimated joint positions in every frame.

Error metrics. We apply the PCKh error metric from [2], which is commonly used for the evaluation of pose estimation approaches [5,44,12,36]. 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. 3, (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. 5 (a), and average PCKh per sequence in Fig. 5 (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. 5 (a)) as well as for some sequences (1, 2, 3, 6, 7, 11, Fig. 5 (b)) are close to perfect (according to the error metric). We observe largest errors when the limbs are directed towards the camera.

OpenPose has reportedly shown impressive results on unconstrained scenes containing adults [5], 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).

4.2 3D Pose Estimation in Depth Images

We evaluate the system with the lowest reported average joint position error for RGB-D data from our overview (Tab. 1), the extension of the pixelwise body part classifier based on random ferns [16,15]. For each pixel of an input depth image, the label is predicted as one of 21 body parts (spine, chest, neck, head, shoulders, elbows, hands, fingers, hips, knees, feet, and toes). The 3D joint positions are then calculated as the mean of each body part region. RGB is not used in this approach. Similar to Sec. 4.1, we calculate joint offsets in the T-pose frame and add them to the estimated joint positions throughout each sequence.

Error metrics. We use the PCKh error metric as described above, here in 3D. Average 3D head segment length over all sequences is 2.64 cm. Additionally, we evaluate the average joint position error (AJPE), which denotes the euclidean distance from estimated joint position to corresponding ground truth.

Results. We present results in Fig. 6. Mean average precision, i.e. average PCKh over all joints and sequences, for PCKh 1.0 is 64.2%, and 90.4% for PCKh 2.0. Compared to the RGB evaluation, we experience a bigger difference between PCKh 1.0 and 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. 6 (a)). 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 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. Joint position errors are largest for the extremities, at an average distance to ground truth of up to 5 cm (Fig. 6 (c)). If the estimate for a joint was missing in a frame, 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

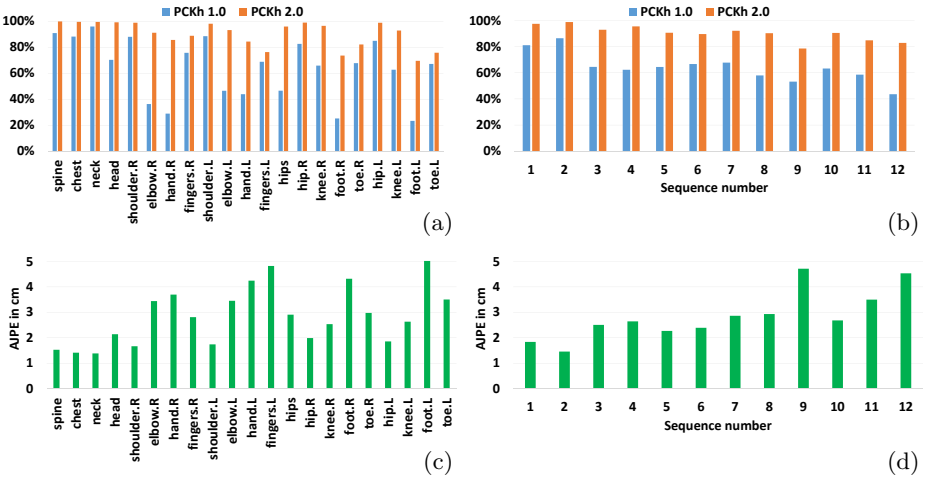


Fig. 6. RGB-D evaluation. Results for 3D pose estimation based on random ferns [16,15]. (a) Percentage of correct keypoints in relation to head segment length (PCKh) per joint. *PCKh 1.0* denotes a correctness threshold of one time head segment length, *PCKh 2.0* of twice the head segment length. (b) PCKh per sequence. (c) Average joint position error (AJPE) per joint. (d) AJPE per sequence.

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. 6 (d)). Best AJPE results are achieved for sequence 2, at 1.46 cm, which is close to results reported in [15]. The varying accuracy for different sequences shows the levels of difficulty and the variance of motion patterns included in the data set.

5 Conclusions

We presented an overview of the state-of-the-art in medical infant motion analysis, with a focus on vision-based approaches and their evaluation. We observed non-standardized evaluation procedures, which we trace back to the lack of publicly available infant data sets. The recently introduced SMIL model allows us to generate realistic RGB and depth images with accurate ground truth 2D and 3D joint positions. We create the Moving INfants In RGB-D (MINI-RGBD) data set, containing 12 sequences of real infant movements with varying realistic textures, shapes and backgrounds. The privacy of recorded infants is preserved by not using real shape and texture, but instead generating new textures and shapes by averaging data from multiple infants. We provide baseline evaluations for RGB and RGB-D data. By releasing the data set, we hope to stimulate research in medical infant motion analysis.

Future work includes the creation of a larger data set, suitable for training CNNs for estimating 3D infant pose from RGB-D data.

References

1. Adde, L., Helbostad, J.L., Jensenius, A.R., Taraldsen, G., Støen, R.: Using computer-based video analysis in the study of fidgety movements. Early human development **85**(9), 541–547 (2009)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3686–3693 (2014)
3. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In: 2015 IEEE International Conference on Computer Vision (ICCV) (2015)
4. Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1302–1310 (2017)
6. Cenci, A., Liciotti, D., Frontoni, E., Zingaretti, P., Carinelli, V.P.: Movements analysis of preterm infants by using depth sensor. International Conference on Internet of Things and Machine Learning (IML 2017) (2017)
7. Chen, H., Xue, M., Mei, Z., Bambang Oetomo, S., Chen, W.: A review of wearable sensor systems for monitoring body movements of neonates. Sensors **16**(12), 2134 (2016)
8. Chen, L., Wei, H., Ferryman, J.: A survey of human motion analysis using depth imagery. Pattern Recognition Letters **34**(15), 1995–2006 (2013)
9. Fan, M., Gravem, D., Cooper, D.M., Patterson, D.J.: Augmenting gesture recognition with erlang-cox models to identify neurological disorders in premature babies. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. pp. 411–420. ACM (2012)
10. Gravem, D., Singh, M., Chen, C., Rich, J., Vaughan, J., Goldberg, K., Waffarn, F., Chou, P., Cooper, D., Reinkensmeyer, D., et al.: Assessment of infant movement with a compact wireless accelerometer system. Journal of Medical Devices **6**(2), 021013 (2012)
11. Hadders-Algra, M., Nieuwendijk, A.W., Maitijn, A., Eykern, L.A.: Assessment of general movements: towards a better understanding of a sensitive method to evaluate brain function in young infants. Developmental Medicine & Child Neurology **39**(2), 88–98 (1997)
12. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: Towards viewpoint invariant 3D human pose estimation. In: European Conference on Computer Vision. pp. 160–177. Springer (2016)
13. Heinze, F., Hesels, K., Breitbach-Faller, N., Schmitz-Rode, T., Disselhorst-Klug, C.: Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy. Medical & biological engineering & computing **48**(8), 765–772 (2010)
14. Hesse, N., Pujades, S., Romero, J., Black, M.J., Bodensteiner, C., Arens, M., Hofmann, U.G., Tacke, U., Hadders-Algra, M., Weinberger, R., Müller-Felber, W., Schroeder, A.S.: Learning an infant body model from rgb-d data for accurate full body motion analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2018)

15. Hesse, N., Schröder, A.S., Müller-Felber, W., Bodensteiner, C., Arens, M., Hofmann, U.G.: Body pose estimation in depth images for infant motion analysis. In: IEEE 39th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC) (2017)
16. Hesse, N., Stachowiak, G., Breuer, T., Arens, M.: Estimating body pose of infants in depth images using random ferns. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015)
17. Iarlori, S., Ferracuti, F., Giantomassi, A., Longhi, S.: RGBD camera monitoring system for alzheimer's disease assessment using recurrent neural networks with parametric bias action recognition. IFAC Proceedings Volumes **47**(3), 3863–3868 (2014)
18. Jiang, C., Lane, C.J., Perkins, E., Schiesel, D., Smith, B.A.: Determining if wearable sensors affect infant leg movement frequency. Developmental neurorehabilitation pp. 1–4 (2017)
19. Kanemaru, N., Watanabe, H., Kihara, H., Nakano, H., Takaya, R., Nakamura, T., Nakano, J., Taga, G., Konishi, Y.: Specific characteristics of spontaneous movements in preterm infants at term age are associated with developmental delays at age 3 years. Developmental Medicine & Child Neurology **55**(8), 713–721 (2013)
20. Karch, D., Kang, K.S., Wochner, K., Philippi, H., Hadders-Algra, M., Pietz, J., Dickhaus, H.: Kinematic assessment of stereotypy in spontaneous movements in infants. Gait & posture **36**(2), 307–311 (2012)
21. Karch, D., Kim, K.S., Wochner, K., Pietz, J., Dickhaus, H., Philippi, H.: Quantification of the segmental kinematics of spontaneous infant movements. Journal of biomechanics **41**(13), 2860–2867 (2008)
22. Kontschieder, P., Dorn, J.F., Morrison, C., Corish, R., Zikic, D., Sellen, A., D'Souza, M., Kamm, C.P., Burggraaff, J., Tewarie, P., et al.: Quantifying progression of multiple sclerosis via classification of depth videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 429–437. Springer (2014)
23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (2015)
24. Loper, M.M., Black, M.J.: OpenDR: An approximate differentiable renderer. In: European Conference on Computer Vision (ECCV). pp. 154–169. Springer (2014)
25. Machireddy, A., van Santen, J., Wilson, J.L., Myers, J., Hadders-Algra, M., Song, X.: A video/IMU hybrid system for movement estimation in infants. In: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 730–733. IEEE (2017)
26. Marcroft, C., Khan, A., Embleton, N.D., Trenell, M., Plötz, T.: Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. Frontiers in neurology **5** (2014)
27. Meinecke, L., Breitbach-Faller, N., Bartz, C., Damen, R., Rau, G., Disselhorst-Klug, C.: Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy. Human movement science **25**(2), 125–144 (2006)
28. Morrison, C., Culmer, P., Mentis, H., Pincus, T.: Vision-based body tracking: turning Kinect into a clinical tool. Disability and Rehabilitation: Assistive Technology **11**(6), 516–520 (2016)
29. Olsen, M.D., Herskind, A., Nielsen, J.B., Paulsen, R.R.: Model-based motion tracking of infants. In: Computer Vision-ECCV 2014 Workshops. pp. 673–685. Springer (2014)

30. Olsen, M.D., Herskindt, A., Nielsen, J.B., Paulsen, R.R.: Body part tracking of infants. In: 22nd International Conference on Pattern Recognition (ICPR). pp. 2167–2172. IEEE (2014)
31. OpenPose library: https://github.com/CMU-Perceptual-Computing-Lab/openpose, accessed June 2018
32. Prechtl, H.F., Einspieler, C., Cioni, G., Bos, A.F., Ferrari, F., Sontheimer, D.: An early marker for neurological deficits after perinatal brain lesions. The Lancet **349**(9062), 1361–1363 (1997)
33. Prechtl, H.: Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. Early human development **23**(3), 151–158 (1990)
34. Rahmati, H., Dragon, R., Aamo, O.M., Adde, L., Stavdahl, Ø., Van Gool, L.: Weakly supervised motion segmentation with particle matching. Computer Vision and Image Understanding **140**, 30–42 (2015)
35. Rihar, A., Mihelj, M., Pašič, J., Kolar, J., Munih, M.: Infant trunk posture and arm movement assessment using pressure mattress, inertial and magnetic measurement units (imus). Journal of neuroengineering and rehabilitation **11**(1), 133 (2014)
36. Sciortino, G., Farinella, G.M., Battiato, S., Leo, M., Distante, C.: On the estimation of children's poses. In: International Conference on Image Analysis and Processing. pp. 410–421. Springer (2017)
37. Serrano, M.M., Chen, Y.P., Howard, A., Vela, P.A.: Lower limb pose estimation for monitoring the kicking patterns of infants. In: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 2157–2160. IEEE (2016)
38. Shivakumar, S.S., Loeb, H., Bogen, D.K., Shofer, F., Bryant, P., Prosser, L., Johnson, M.J.: Stereo 3D tracking of infants in natural play conditions. In: International Conference on Rehabilitation Robotics (ICORR). pp. 841–846. IEEE (2017)
39. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4645–4653 (2017)
40. Singh, M., Patterson, D.J.: Involuntary gesture recognition for predicting cerebral palsy in high-risk infants. In: International Symposium on Wearable Computers (ISWC). pp. 1–8. IEEE (2010)
41. Spittle, A., Orton, J., Anderson, P.J., Boyd, R., Doyle, L.W.: Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. The Cochrane Library (2015)
42. Stahl, A., Schellewald, C., Stavdahl, Ø., Aamo, O.M., Adde, L., Kirkerød, H.: An optical flow-based method to predict infantile cerebral palsy. IEEE Transactions on Neural Systems and Rehabilitation Engineering **20**(4), 605–614 (2012)
43. Sun, B., Liu, X., Wu, X., Wang, H.: Human gait modeling and gait analysis based on Kinect. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 3173–3178. IEEE (2014)
44. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4732 (2016)
45. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3D human pose estimation in RGBD images for robotic task learning. In: IEEE International Conference on Robotics and Automation (ICRA) (2018)