Concurrent Validity of a Custom Method for Markerless 3D Full-Body Motion Tracking of Children and Young Adults Based on a Single RGB-D Camera

Hesse, Nikolas ; Baumgartner, Sandra ; Gut, Anja ; van Hedel, Hubertus J A

DOI: https://doi.org/10.1109/TNSRE.2023.3251440

Concurrent Validity of a Custom Method for Markerless 3D Full-Body Motion Tracking of Children and Young Adults Based on a Single RGB-D Camera

Nikolas Hesse, Sandra Baumgartner, Anja Gut, and Hubertus J. A. van Hedel

***Abstract*— Low-cost, portable RGB-D cameras with integrated body tracking functionality enable easy-to-use 3D motion analysis without requiring expensive facilities and specialized personnel. However, the accuracy of existing systems is insufficient for most clinical applications. In this study, we investigated the concurrent validity of our custom tracking method based on RGB-D images with respect to a gold-standard marker-based system. Additionally, we analyzed the validity of the publicly available Microsoft Azure Kinect Body Tracking (K4ABT). We recorded 23 typically developing children and healthy young adults (aged 5 to 29 years) performing five different movement tasks using a Microsoft Azure Kinect RGB-D camera and a marker-based multi-camera Vicon system simultaneously. Our method achieved a mean per joint position error over all joints of 11.7 mm compared to the Vicon system, and 98.4% of the estimated joint positions had an error of less than 50 mm. Pearson's correlation coefficients *r* ranged from strong (r =0.64) to almost perfect (r>0.99). K4ABT demonstrated satisfactory accuracy most of the time but showed short periods of tracking failures in nearly two-thirds of all sequences limiting its use for clinical motion analysis. In conclusion, our tracking method highly agrees with the gold standard system. It paves the way towards a low-cost, easy-to-use, portable 3D motion analysis system for children and young adults.**

***Index Terms*— Children, Kinect, 3D motion tracking, RGB-D, Vicon.**

I. Introduction

THE gold standard technology to perform clinical three-dimensional motion analysis (3DMA) is marker-based multi-camera systems, which track the 3D positions of reflective markers with highest accuracy [1] and transfer their movements onto a biomechanical model. In combination with force plates and electromyographic recordings, 3DMA is routinely applied to collect spatiotemporal, kinematic, kinetic, and electromyographic data for gait analysis. These data are needed to identify and understand the underlying impairments and support the management of gait deviations in children with cerebral palsy [2]. Some studies have demonstrated benefits of 3DMA also for other motor tasks [3] like upper limb function [5, 4] or trunk control [6].

However, marker-based multi-camera systems are bulky, non-portable, costly, and require regular calibration before use. In preparation for 3DMA recordings, trained personnel must precisely place reflective markers on predefined anatomical locations while patients should cooperate and keep still. This procedure can be challenging for therapists and patients, especially for children with cognitive impairments, who might not tolerate such lengthy procedures. Furthermore, post-processing requires time, e.g., filling gaps in marker trajectories in case of occlusions. Finally, a high level of expertise is needed to interpret the findings. These requirements have precluded the adoption of 3DMA for routinely applied assessments of motor function, except for 3D gait analysis, as mentioned above.

RGB-D cameras provide a simpler, inexpensive alternative for capturing motion in 3D. Their integrated body tracking functionality produces estimates of 3D body joint positions in each video frame. Extracted kinematics from sensors like the Microsoft Kinect v1 and v2 or the recently released Azure Kinect Developer Kit (AKDK) have been evaluated for clinical motion analysis [7], [8], [9]. While some spatiotemporal variables, like gait speed, step length, and stride time, agree highly with those derived from gold standard systems [10, 11, 12], the accuracy of the kinematics obtained with the Kinect v1 and v2 appears to be limited [13], [14], [15], with questionable validity and reliability [16] and too inconsistent for clinical use [10]. Although direct comparisons between the body tracking software associated with the more recent depth sensor AKDK (termed Azure Kinect Body Tracking, or, in short, K4ABT) and gold standard 3DMA systems have been performed [8], [17], these results have to be interpreted with caution. The reflective markers of the reference systems distort the depth data and lead to more frequent tracking

Manuscript received 6 July 2022; revised 22 November 2022 and 1 February 2023; accepted 24 February 2023. Date of publication 2 March 2023; date of current version 10 April 2023. This work was supported by the Anna Mueller Grocholski Foundation. *(Corresponding author: Nikolas Hesse.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Committee of the Canton of Zurich, Switzerland, under Application No. BASEC-Nr. PB_2016-01843.

The authors are with Swiss Children's Rehab, University Children's Hospital Zurich, 8910 Affoltern am Albis, Switzerland, and also with the Children's Research Center, University Children's Hospital Zurich, University of Zurich, 8032 Zurich, Switzerland (e-mail: Nikolas.hesse@kispi.uzh.ch; sandra.baumgartner@kispi.uzh.ch; anja.gut@kispi.uzh.ch; Hubertus.vanhedel@kispi.uzh.ch).

Code will be made available at https://github.com/nh236/smplify-kids

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3251440, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2023.3251440

failures [9]. While Yeung et al. reported superior body tracking performance of K4ABT compared to Kinect V2 during gait, it should be noted that they did not compare the same gait cycles but instead recorded leg joint angles during different bouts of treadmill walking [9].

We are therefore convinced that the tracking methods of the Kinect v1 and v2 do not provide the accuracy and reliability required for clinical full-body motion analysis, and that further investigation of K4ABT is necessary. This is especially true for children, for whom the accuracy of approaches trained primarily on adult data generally decreases [18].

RGB-D cameras provide body tracking capabilities and grant access to raw color (RGB) and depth (D) images, thus allowing the development of custom 3D body tracking methods. However, while motion tracking from RGB-D data is an active area of research [19], [20], [21], [22], current methods cannot track the motions of (smaller) children. In addition, these methods have not been compared to a gold standard system.

In this study, we adapted an existing method that was developed to track infants from RGB-D sequences [22] to work with humans of all sizes, i.e., for children and adults, and evaluated its concurrent validity with respect to a gold standard system.

Our first aim was to determine the accuracy of our method against a marker-based Vicon system on a data set derived from typically developing children (TDC) and young adults performing five tasks. Our second aim was to determine the validity of tracking results of the publicly available method K4ABT in the same group of children and young adults.

II. Method

A. Participants

We included TDC (from the age of 5 years) and young adults (<30 years) utilizing purposive sampling to cover a broad range of ages and body sizes. The TDC were children of employees of the Swiss Children's Rehab and colleagues. The young adults worked at the rehabilitation center. Participants were excluded if they had a neurological, musculoskeletal, or cardiovascular diagnosis. All participants, and the parents of participants aged less than 18 years, were informed verbally and in writing about the study. All participants had to provide verbal agreement to participate. Children aged 14 years and older and adults also had to provide written informed consent. Each participant could withdraw from the study at any time. The Ethical Committee of the Canton of Zurich approved the study (BASEC-Nr. PB_2016-01843).

B. Movement Tasks

We selected five tasks that we adopted from clinical assessments of motor performance, e.g., from assessing trunk control [23] or gross motor functions [24], or pose challenges for tracking systems due to (self-)occlusions or high movement velocity. Each task was explained verbally and demonstrated once. The starting position for all tasks was upright standing while facing the camera with arms hanging down. The tasks were performed in the same order and defined as follows:

- *Reach to the other side*. Participant reaches with one hand across the body to the other side, roughly at shoulder height, then repeats the task with the other hand.
- *Trunk bending*. Participant bends the trunk to the left, the right, forward, and backward.
- *Standing straight leg raise*. The participant raises one straight leg to the front, holds for three seconds, and returns to the starting position, then repeats the task with the other leg. We chose this task despite the fact that most children with motor impairments would not be able to execute it, because it contains a balance component and the raised leg causes self-occlusions, which makes the motion more difficult to track.
- *Squats*. The participant performs three squats.
- *Jumping jacks*. The participant performs jumping jacks three times. While most children with motor impairments cannot perform this task, we included it because the fast motions are challenging to track.

The experiments took place in our gait lab. Preparing the participants was time-consuming, i.e., placing the reflective markers, performing the body measurements, and collecting patient information (30 minutes). In contrast, executing the five motor tasks took less than 5 minutes. We recorded the participants with an AKDK RGB-D camera and a marker-based Vicon system simultaneously. At the end of each session, we removed the markers, and the participants repeated the five tasks, this time recorded only with the AKDK.

III. Data Processing

A. Setup

The Vicon system consisted of 12 Vero V2.2 cameras to capture the marker set of the Conventional Gait Model (CGM) 2.5 [25], which consists of 51 markers. More detailed marker models for the upper body exist, but the CGM provided a good trade-off between accuracy and usability for our evaluation, including young children. We recorded data at 120 Hz, except for two participants, who were recorded at 90 Hz. The diameter of the markers was 16 mm. The marker data was post-processed with Vicon Nexus 2.10 for filling small gaps in marker trajectories and calculating joint centers from the marker positions.

The AKDK camera was mounted on a tripod, facing the participant frontally at a distance at which the entire body was visible (between 1.5 and 2.9 meters). We recorded in "NFOV unbinned" mode at a depth resolution of 640×576 and color image resolution of 1920×1080. The depth and RGB images were registered, and 3D point clouds were computed using the camera calibration, using methods from the Microsoft Sensor SDK [26].

B. Interference Between AKDK and Vicon

The AKDK relies on the Time-of-Flight principle, where the camera emits infrared light pulses and measures the time it takes the light to bounce back from the scene to the sensor [27]. The active illumination of the Vicon system, which also operates in the infrared spectrum, interfered with the AKDK depth measurements. To avoid this effect and at the same time acquire temporally aligned measurements with the AKDK, we used the Vicon's synchronization pulse and

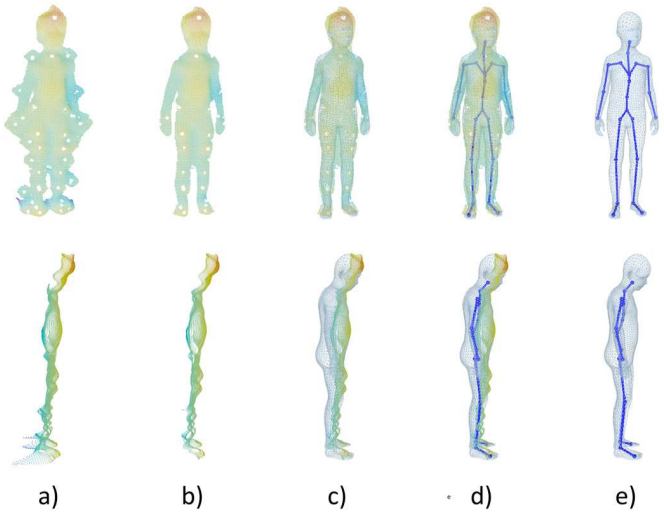

Fig. 1. 3D point cloud captured with AKDK and tracking results. For anonymization, we do not color the point cloud with the RGB information, but instead use a color coding where points close to the camera are red, and points further away blue. Top: frontal view, bottom: side view. a) Point cloud of standing child with attached reflective markers. Note how the markers cause missing points (small white dots) and spherical bumps. b) Processed point cloud with reduced marker noise based on segmentation of color image. Noise besides the body is removed, but not the bumps on the frontal side of the body. c) Point cloud with aligned body model (tracking result), d) model with underlying skeleton, and e) model and skeleton without point cloud.

added an offset of less than one millisecond. The AKDK recorded 30 frames per second, and we selected the Vicon measurements with timestamps closest to those of the AKDK to get joint positions at the same frame rate and account for dropped frames.

As reported in previous literature, even without the active illumination of the Vicon system, the strong reflectivity of the markers creates significant amounts of noise in the depth data solely from the illumination of the AKDK [9]. This noise manifested itself as holes in the 3D point cloud at the position of the markers, surrounded by spherical disturbances, as shown in (Fig. 1, a). To make the data more similar to the actual (markerless) use case, we reduced the quantity of noise at the sides of the body by applying a method for semantic segmentation, deeplabv3 [28], to identify the person's silhouette in the color image, and crop the depth image to the segmentation mask (Fig. 1, b). We applied this to all except the jumping jacks sequences since the motion blur in the color images led to poor segmentation results.

C. Our Method

Our tracking method is based on a parametric model of the 3D body surface, termed *SMPL-H* [29]. Such a virtual body aims at simulating real humans in that it can realistically vary its *shape*, i.e., body characteristics like height and volume, and *pose*, i.e., joint angles of an underlying skeleton.

The input to our method is a 3D point cloud computed from an RGB-D image, and the output is an SMPL-H body that matches the body in the 3D point cloud (Fig. 1, c-e). From this virtual body, kinematic variables like joint positions and angles can be extracted as a proxy for the real body.

To obtain the shape and pose parameters of the parametric body model from RGB-D data, we adapted the method of Hesse et al. [22], which estimated the pose and shape of infants from sequences of 3D point clouds. The pose and shape parameters of the infant body model, termed *SMIL*, are automatically adjusted by iteratively minimizing an objective function until the model surface matches that of the body in the point cloud, as is standard practice [20], [22], [30].

Both models SMIL and SMPL-H are not able to properly represent bodies outside the learned populations, i.e., infants and adults (see Supplementary material). Our main modification of the method by Hesse et al. is the extension of the body model to work for humans of all sizes by adding a parameter α that determines the interpolation between infant and adult shapes, similar to Patel et al. [31]. The parameter α lies in the range of 0 and 1, with 0 representing an infant shape, and 1 an adult shape, and is automatically adapted during the optimization, particularly in the initialization stage.

In the original SMIL model, the hands are shaped as fists, while the adult SMPL-H model has open hands. This leads to distorted hands during interpolation, making hand/finger tracking impossible. For this reason, we created a new infant template with open hands. We extracted an infant mesh with open hands from the open-source software MakeHuman [32], transferred it to SMIL topology similar to [22], and replaced the hand vertices of the SMIL template with those of the new mesh. This allows us to track finger motions.

We optimize the objective function for the shape and pose parameters, β and θ, and the interpolation parameter α:

$$\mathrm{E}(\beta, \theta, \alpha) = E_{\text{data}} + E_{\text{prior}} + E_{\text{temporal}}$$

E_{data} consists of the point-to-plane distances between the 3D point cloud and the model, and the distance between estimated 2D joint positions (body and hand) and model joints projected to 2D. It includes a penalty for model points penetrating the (estimated) ground plane and a gravity term to pull model points near the ground plane toward it. The computed distances are processed with a robust Geman-McClure error function [33] to avoid fitting to noise/outliers. We additionally use an interpenetration term, similar to [30], to penalize self-intersections. E_{prior} contains a regularization term to keep pose and shape plausible. Instead of using the infant pose prior from [22], we used one trained on adult data from [34]. E_{temporal} supports temporal smoothness by keeping the model close to the fitting result of the previous frame. Each of the terms in the objective function has an associated weight that we determined experimentally. Note that while each of the terms influences the result, the weights are chosen to not restrain the model from completely obeying one of the constraints. If the data term shows clear evidence of where the model should go, this will likely overrule the prior term, e.g., in the case of "unhealthy" poses.

During initialization, we run multiple rounds of fitting, starting with high weights for the pose and shape prior and lower weights for the data terms, and moving towards high weights for data terms and lower weights for the priors. More details can be found in the supplementary material.

For the optimization of the objective function, we use an LBFGS optimizer with strong Wolfe line search [35]. Our implementation is based on the publicly available code of SMPLify-X [30] and is implemented in Python using the

packages pytorch and pytorch3d [36]. We performed the processing on a Desktop PC with a NVIDIA GeForce RTX 2080 GPU with 8 GB RAM. Fitting took between 0.5 and 5 seconds per frame. We used the male version of SMPL-H for children and male adults, and the female version for female adult participants.

D. K4ABT

We extracted 3D joint positions from the same RGB-D sequences that were input to our method using the publicly available *Azure Kinect Body Tracking SDK* (K4ABT), version 1.1.2, with standard settings [37].

E. Joints

For our analysis, we selected body joints from SMPL-H and K4ABT that correspond to joint centers derived from the Vicon system, namely the head, shoulders, elbows, wrists, hips, knees, ankles, and feet. We transformed the body joint positions from the AKDK coordinate frame and the Vicon system to the same reference coordinate frame. We found that skeleton definitions differ between SMPL-H and Vicon, i.e., that joint centers are defined in different locations with respect to the body surface. For this reason, we did not apply a rigid transformation to the entire skeleton but to each joint separately. The intuition behind this was that we intend to analyze *movements*, i.e., how a point moves over time instead of its global position in space. We assume that points close to each other move similarly, allowing us to compensate for differences in skeleton definition by bringing joint trajectories into correspondence. We used *Singular Value Decomposition* (SVD) to find the rotation and translation to map between two sets of corresponding joint positions [38]. However, SVD is sensitive to outliers and noise. Therefore, we implemented a *RANSAC* scheme [39] by repeatedly sampling a third of the joint positions at random. We found the transformation using SVD and selected the rotation and translation with the highest percentage of inliers (distance between pairs of corresponding joints < 10 mm). For a comparison of sequences captured *without* markers, the transformation from K4ABT joints to the SMPL-H skeleton was calculated similarly. We did not apply smoothing to any of the joint positions. All evaluations are based on joint positions that were transformed with this procedure.

F. Evaluation Metrics and Statistics

We evaluated multiple metrics because a single metric cannot capture all relevant aspects of tracking accuracy for motion analysis.

We computed the *mean per joint position error* (MPJPE), i.e., the mean Euclidean distance between Vicon joint positions and those predicted by our method. MPJPE calculates the mean value of the complete sequence, which is why sporadic tracking failures, i.e., large differences between estimated and reference joint positions, only have a relatively small effect on the MPJPE value but may negatively impact motion analysis.

Hence, we computed the *percentage of correct keypoints* (PCK), which is the fraction of estimated joint positions within a distance to the reference system that is smaller than a threshold τ, which we chose to range from 5 to 200 mm.

The agreement of the movement signals over time is an important property for motion analysis, which we evaluate using *Pearson's correlation coefficient* (r). A small amount of noise is inherent to depth data, which is why correlations will be low in cases where joints are stationary, e.g., in the lower limbs during *trunk bending* or *reaching*. Instead of presenting the average r-value over all joints, we examined the influence of movement magnitude by evaluating r with respect to the amount of motion present in a joint, which we represented using the *standard deviation* (SD) over joint positions in a sequence, arranged in 5 categories. For each joint in each sequence, the SD value was calculated from the reference system in 3 dimensions (X: medio-lateral, Y: vertical, and Z: anterior-posterior). Then, the r value of each dimension was assigned to a category according to the SD value. The total number of items that were categorized is *#sequences* $\times$ *#joints* $\times$ *#dimensions*. Finally, we calculated the average of all r-values in each category per dimension. We additionally computed the percentage of items per category to show the distribution of joints with different movement magnitudes. For better interpretability of the SD metric, some exemplary SD values for a five-year-old child during the reach task: the average SD over lower limbs, which were standing still during the whole sequence, computed from the Vicon system was 5, 1, and 5 mm in X, Y, and Z dimension, respectively. The average SD in upper limbs was 70, 58, and 69 mm for X, Y, and Z dimensions, with a maximum for the wrist joints at 126, 111, and 89 mm.

As mentioned above, Vicon markers interfere with the AKDK recordings. This not only influences the depth images but heavily affects the tracking quality of K4ABT, leading to repeated tracking failures when markers are present on a person's body [9]. Hence, we refrained from a direct comparison between K4ABT and Vicon since this does not reflect the actual capabilities of K4ABT. Without the Vicon system as a reference, the question remained how to reasonably analyze the K4ABT tracking results. The lack of ground truth prevented a detailed evaluation of the accuracy. Therefore, we focused on significant tracking errors. First, we analyzed the percentage of complete tracking failures for recordings *with* and *without* markers, i.e., the fraction of frames in which K4ABT did not detect a body. This was a fair comparison, as no reference system was required. Second, we used our method as reference system to evaluate K4ABT on the sequences recorded *without* markers (and the Vicon system). We restricted our evaluation to metrics that corresponded (almost) perfectly to the Vicon system. To avoid biasing the results to the disadvantage of K4ABT, we visually verified the correctness of our method's predictions, and excluded few cases in which our method did not produce satisfactory results from the evaluation of K4ABT.

IV. Results

Eighteen TDC and five young adults (14 females, 9 males) aged 5 to 29 years (mean: 13.2 years), with a body height between 110 and 189 cm (mean 148.4 cm), body weight between 19 and 77 kg (mean 43.7 kg), and BMI between

13.8 and 26.0 (mean 18.9) participated. Each participant performed the five movement tasks, *with* and *without* markers, giving a total of 115 recordings per condition. Due to technical problems, we had to exclude eight recordings *with* markers from three participants, resulting in 107 recordings. For two subjects, no recordings *without* markers were performed, leaving 105 recordings *without* markers. The movement tasks are displayed in Fig. 2 and the supplementary video.

A. Evaluation of Our Method

We present results for the MPJPE per motion task in Table I and the PCK, i.e., the percentage of estimated body joints with errors below the threshold τ, in Table II. In Table III, we display Pearson's correlation coefficients r for three dimensions, categorized for the amount of motion per joint and sequence.

Average MPJPE values per joint ranged from 8.8 mm to 15.5 mm, at an average of 11.7 mm over all joints and all motion tasks. Regarding the PCK metric, 63.8% of the estimated joint positions were within 10 mm of the gold standard joint positions, 98.5% of joints were assessed with an error of less than 50 mm, and hardly any estimations exceeded the error limit of 100 mm. The more motion was present in a joint, the higher r was, ranging from 0.5-0.7 for joints that were not moving at all (SD $<$ 10 mm), over 0.95 for joints that were slightly moving (SD 10-30 mm) to near perfect correspondence $>$ 0.99 for joints that moved a lot (SD $>$ 100 mm). All but a few items of the stationary group had p-values $<$ 0.0001.

During *leg raise* and *squats*, we observed slightly increased MPJPE values for the hip and knee, and elbow and shoulder, respectively. These occurred when limbs pointed directly at the camera, and the foot or hand occluded most of the leg or arm, leading to small inaccuracies, as seen in Fig. 2.

The task containing very fast movements of all body parts, *jumping jacks*, showed a comparatively high average MPJPE of 17 mm over all joints. The fast motions led to missing points in the depth image (cf. Fig. 2, bottom row, and the supplementary video), making pose detection very difficult. Despite the flawed data, we experienced very few inaccurate results (0.025 % of joint positions with error > 100 mm, see Table II, *column jump*), concerning exclusively wrist joints in 4 of 23 sequences, and no complete tracking failures (PCK $=$ 100% for $\tau = 200$ mm). The MPJPE for feet is high because the noise introduced by markers sometimes led to an outside foot rotation during landing to match the 3D points. The lack of stationary joints is illustrated by the low PCK value for errors smaller than 5 mm (8.2%). As mentioned previously, the more motion was present in a joint, the higher its correlation with the gold standard system, with r values exceeding 0.99 for SD values of more than 100 mm (see Table III, last row).

B. Effect of Markers on K4ABT

K4ABT did not detect any body at all in 44 of 107 sequences *with* markers. For the *reach* task, 10 sequences were affected, of which 1.7% to 64.1% of the frames were without detected body. Seven of the *trunk* sequences (0.9%-48.8% of frames), two of the *leg raise* sequences (16.7%-30% of frames), 8 of the *squats* sequences (1.5%-58.3% of frames),

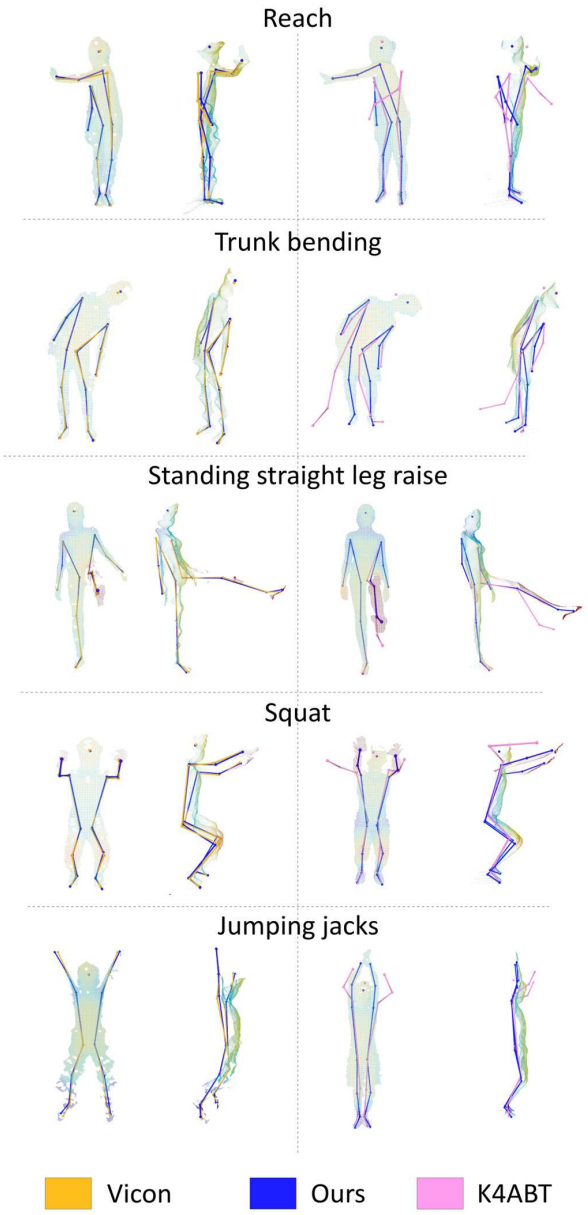

Fig. 2. 3D point clouds with estimated joint positions for different motion tasks. Yellow: Vicon, blue: our method, pink: K4ABT. Front and side view. Left: Vicon vs our method (recordings with markers). Right: Our method vs K4ABT (recordings without markers). Images for same task (in same row) are taken from sequences of the same participant.

and 17 of the *jumping jacks* sequences (0.6%-79.9% of frames) were affected. The failures mainly occurred in sequences of children. In contrast, sequences of adults only showed sporadic failures (overall in 3 sequences: one *trunk* sequence with 2.3% of frames and two *jumping jacks* sequences with 0.6% and 1.4% of frames without detected body). For sequences *without* markers, only 11 of 105 sequences contained frames in which no body was detected. Apart from one *reach* sequence with 0.6% of missing detections, these failures appeared exclusively in the *jumping jacks* task, affecting between 0.5% and 11.1% of frames in the sequences of mostly younger children.

C. Evaluation of K4ABT

The influence of markers on K4ABT tracking quality was reflected in the MPJPE of K4ABT with respect to our

TABLE I
MEAN PER JOINT POSITION ERROR (MPJPE) PER TASK

Joint	Reach	Trunk	Leg raise	Squats	Jump	Average
Head	9.1 (4.7)	11.4 (4.2)	6.4 (2.1)	8.2 (2.7)	10.9 (3.1)	9.2 (1.8)
Shoulder	19.4 (5.0)	11.3 (3.4)	6.0 (2.1)	11.9 (5.2)	19.1 (3.2)	13.6 (5.1)
Elbow	20.3 (7.8)	13.0 (4.2)	8.6 (3.3)	17.3 (9.8)	18.5 (3.8)	15.5 (4.2)
Wrist	21.7 (14.1)	9.4 (4.7)	8.5 (4.7)	13.9 (7.4)	21.4 (5.5)	15.0 (5.7)
Hip	6.9 (2.2)	5.7 (1.6)	12.1 (4.5)	9.8 (3.5)	9.3 (2.1)	8.8 (2.3)
Knee	4.8 (1.5)	5.2 (1.4)	7.7 (3.4)	12.2 (4.0)	14.5 (1.9)	8.9 (3.9)
Ankle	4.8 (1.2)	6.2 (1.7)	8.1 (2.4)	12.0 (5.7)	15.5 (2.1)	9.3 (3.9)
Foot	5.4 (2.4)	6.2 (1.8)	15.0 (4.8)	11.8 (5.1)	24.3 (5.3)	12.5 (6.9)
Average	11.7 (2.9)	8.3 (1.6)	9.3 (2.1)	12.4 (4.7)	17.0 (1.4)	11.7 (3.0)

MPJPE of our method with respect to the Vicon system, in mm (standard deviation in brackets). Results for left and right sides were averaged.

TABLE II
PERCENTAGE OF CORRECT KEYPOINTS (PCK) PER TASK

τ in mm	Reach	Trunk	Leg raise	Squats	Jump	Average
5	33.745	34.281	30.129	19.474	8.232	25.172
10	69.759	77.392	74.157	59.474	38.266	63.810
20	85.041	93.702	91.323	84.058	69.124	84.650
30	92.065	98.243	96.721	92.906	85.783	93.143
40	95.377	99.414	98.940	96.543	93.618	96.778
50	97.281	99.776	99.608	98.080	97.622	98.473
100	99.490	99.971	99.969	99.925	99.975	99.866
200	99.860	100.000	99.997	99.998	100.000	99.971

Percentages of estimated joint positions by our method with error smaller than threshold τ, in %.

TABLE III
AVERAGE PEARSON'S CORRELATION COEFFICIENTS *r*

Amount of motion (SD in mm)	r_X	r_Y	r_Z	% items
<10	0.757 (0.22)	0.624 (0.27)	0.532 (0.33)	24.1
10-30	0.957 (0.07)	0.942 (0.08)	0.957 (0.06)	20.0
30-50	0.988 (0.02)	0.976 (0.03)	0.975 (0.04)	14.9
50-100	0.993 (0.02)	0.991 (0.01)	0.990 (0.02)	24.1
>100	0.993 (0.03)	0.998 (0.00)	0.996 (0.02)	16.9

Correlation of joint positions over time between our method and the Vicon system, with respect to the amount of motion present per joint per sequence. Values for *r* (standard deviation in brackets) are displayed per axes X (medio-lateral), Y (vertical), and Z (anterior-posterior). The amount of motion is computed per joint per sequence, given by standard deviation (SD) of joint positions, and correlations are grouped according to this value. Groups range from stationary joints (SD < 10 mm) to joints with lots of motion (SD > 100 mm). In the last column, we present the fraction of items included in each group, where one item denotes one dimension of one joint in one sequence. *P*-values were < 0.0001 for all but a few items of the stationary group.

method, at an average MPJPE over all joints of 45.7 mm for recordings *with* markers. We verified this result with Vicon as a reference at an average MPJPE over all joints of 45.1 mm for K4ABT. For recordings *without* markers and our method as the reference, the average MPJPE over all joints for K4ABT was 26.9 mm.

Regarding tracking failures, we also evaluated PCK values for which our method yielded close to perfect results. For sequences *with* markers, K4ABT achieved a PCK for τ = 200 mm of 96.4% with our method as reference (same value with Vicon as reference) and a PCK for τ = 100 mm of 89.8% (89.9% with Vicon as reference). *Without* markers and with our method as the reference, the PCK for τ = 200 mm was 98.2% and 95.5% for τ = 100 mm.

Considering the above results, we concentrated on the recordings *without* markers for the evaluation of K4ABT and used our method as the reference for those metrics in which our method reaches close-to-perfect agreement with the Vicon system. Additionally, we visually verified the results of our method and excluded a few failure cases from the evaluation to not influence the results to the disadvantage of K4ABT.

Regarding the PCK metric, 95.5% of the K4ABT joint positions were estimated within 100 mm of ours (i.e., 4.5% > 100 mm), and 98.2% with errors smaller than 200 mm, meaning that 1.8% of the estimated joints were failure cases. These were distributed across 67 of 105 sequences (19 *reach*, 6 *trunk*, 13 *leg raise*, 12 *squats*, 17 *jumping jacks*). We display examples of failure cases in Fig. 2, right. In the *reach* task, the arm was repeatedly lost when the arm crossed the body midline (PCK with τ = 200 mm: 97.7% for elbows, 94.9% for wrists). K4ABT was relatively stable during *trunk bending*, and we observed only few erroneous joint positions. Errors occurred more commonly when the limbs pointed directly at

the camera, e.g., legs during *leg raise* (PCK with $\tau = 200$ mm: 96.5% for knees, 86.2% for ankles, 87.0% for feet) and arms during *squats* (PCK with $\tau = 200$ mm: 94.8% for elbows, 91.8% for wrists). In the *jumping jacks* sequences, short series of errors frequently occurred when the arms were lifted above the head, or the fast movements caused complete tracking failures at times without detecting any body at all. We did not include frames without K4ABT body detections in the calculation of any error metric.

Pearson's correlation coefficient r for joint positions with a large amount of motion (SD $>$ 100 mm) was very high, with an average value over all dimensions of 0.95 (SD 0.1).

In summary, reflective markers negatively impacted K4ABT, but movement signals of K4ABT and our method agreed most of the time. Nevertheless, K4ABT demonstrated repeated tracking failures of short periods in nearly two-thirds of all sequences.

V. Discussion and Limitations

The primary aim of this study was to evaluate the concurrent validity of a custom method for estimating the 3D pose of children and young adults from RGB-D data with respect to a gold standard system. While our method achieved accurate results, it is not intended to replace specific applications like the comprehensive routine gait analysis, which is required to make clinical decisions, e.g., whether the child should undergo surgery. Instead, we see our method as an objective and more accurate alternative or supplement to standard motor function assessments, i.e., assessments based on the subjective perception of a therapist rating the children's performance on a coarse, relatively unresponsive, ordinal scale.

A. Results – Our Method

We found an average MPJPE of 11.7 mm, a PCK of over 98% for an error threshold of 50 mm, and close to perfect correlation in all dimensions if at least some motion was present in a joint (r values of $\sim$0.95 (SD $>$ 10mm), and $>$ 0.99 if a lot of motion was present (SD $>$ 100mm)). The combination of these error metrics shows that our method can produce valid pose estimates approaching the results of a gold-standard system. Some failures occurred when motion blur caused the segmentation, which we applied to reduce the influence of markers, to fail. However, as the segmentation was only applied to reduce marker-induced noise, this is irrelevant for our target application. Furthermore, while our method seems generally robust to self-occlusions, e.g., hands occluding the face or limbs pointing directly towards the camera occluding the rest of the extremity, rare failures were attributed to limbs being hidden entirely behind the body. This occurred, for example, during the *reach* task where the non-reaching arm disappeared behind the body when a child turned the trunk into the reaching direction or during the *leg raise* task when the trunk covered an arm during balancing. To handle these failures, occlusions can be detected by integrating a measure of tracking confidence proportional to the number of 3D points in close proximity to each body part, which would be very low in the case of hidden limbs. This would allow for an automated exclusion of these cases from subsequent motion analysis.

Our method could be extended to use data from multiple cameras to resolve occlusions, which is an advantage of existing commercial markerless video-based multi-camera systems, e.g., Theia markerless [40], which has been validated for adults during treadmill gait with respect to a marker-based system [41]. This system's average 3D joint position errors range from 11 mm (wrists) to 36 mm (hips). The multi-camera setup reduces occlusions, thus allowing more unconstrained movements, but at the same time, increases space requirements and effort for camera placement and calibration.

B. Related Work – Shape and Pose Estimation From RGB-D

Other methods have been proposed for tracking the full body shape and pose from RGB-D data. Bashirov et al. trained a neural network to predict the shape and pose of a parametric body model from the estimated 3D skeleton of K4ABT [42]. The average positional error of the method is reported to lie in the range of 4 cm. However, this method relies on the output of K4ABT and therefore is subject to its limitations. More similar to our work, Bogo et al. used an optimization-based method to align a parametric body model to 3D point clouds from RGB-D sequences to extract accurate pose, shape, and appearance, from which they created a textured avatar [20]. The focus of this work was the precise reconstruction of body shape and appearance, and no quantitative evaluation of pose accuracy was presented. Rempe et al. introduced an optimization-based method that integrates a learned dynamic motion model to predict the body shape and pose from RGB-D sequences [43]. This approach is powerful in predicting plausible poses under the influence of noise and occlusions of the body by objects. In clinical applications, however, the focus lies on capturing fine-grained changes in movements instead of generating smooth and plausible poses. It is questionable if the motion model, which was trained on data from healthy adults, could generalize to motions of children with motor disorders without correcting the movements to look "healthy". In summary, each of the methods targets tracking adults, and none was validated with respect to a marker-based gold standard system.

C. K4ABT Results

Our second aim was to determine the validity of K4ABT. The negative effect of reflective markers on K4ABT tracking quality has been described previously [9], but a quantitative evaluation was not conducted due to the lack of a markerless reference system. Our study approximated the amount of disturbance of the markers to the accuracy of K4ABT. The number of frames in which no body was detected was much higher *with* markers than *without* markers. Missing body detections occurred in 44 of 107 sequences *with* markers, affecting up to 80% of the frames, and were not restricted to specific tasks. Only 11 of 105 sequences *without* markers were affected. Up to 11% of the frames were without detected body. This occurred in the *jumping jacks* task and once in a *reach* sequence. Particularly smaller children were affected, suggesting that K4ABT is tuned towards adults and has some difficulties recognizing children (*with* and *without* markers).

Using our method as an imperfect reference system, the average MPJPE over all joints and motions for K4ABT nearly doubled from 26.9 mm *without* markers to 45.7 mm *with* markers. Similarly, the percentage of correct keypoints decreased by 2% for $\tau = 200$ mm and approximately 6% for $\tau = 100$ mm for sequences *with* compared to *without* markers.

D. Related Work – K4ABT Validation

Despite these issues, previous studies compared the accuracy of K4ABT to marker-based Vicon systems. Albert et al. analyzed K4ABT tracking accuracy during treadmill walking [8]. They reported MPJPE values of approximately 10 – 15 mm for body parts related to the trunk and head and larger errors for the arms (elbows ∼17 mm, wrists ∼25 mm, hands ∼50 mm) and legs (knees ∼30 mm, ankles ∼60 mm, feet ∼60 mm). These results are similar to our findings *with* markers, taking into account that our motion tasks are more challenging to track than treadmill walking. In another study, K4ABT was validated during lateral and forward reach tasks and single stance balance with eyes closed [17]. The results at MPJPE of 70 mm, 90 mm, and 47 mm, respectively, align with our results *with* markers. The authors concluded that K4ABT showed very high tracking accuracy but low tracking quality for fast movements and a repeated loss of tracking movements along the focal axis [17].

Based on our evaluation of marker influence on K4ABT, the general tracking quality of K4ABT *without* markers is better than reported in studies evaluating it *with* markers. In our study *without* markers, we observed that K4ABT worked quite accurately most of the time. However, we encountered short sections of severe tracking failures in many sequences, with joint positions being relatively far from the person's point cloud. Failure cases in *our method* predominantly involved limbs hidden behind the body not directly involved in the task. In *K4ABT*, however, failure cases included body parts directly involved in the tasks, e.g., the reaching arm during *reach*, the raised leg during *leg raise*, or the knees during *squats* (see Fig. 2 and the supplementary video). While K4ABT can still be a valuable tool, this limits its usability for clinical motion, as tasks should be limited to those that are relatively easy to track, while tracking output should be verified before further processing. Most common failure cases for K4ABT occurred when the arms were moved across the body midline (*reach*), when limbs were directed towards the camera (*squats, leg raise*), or when arms were lifted above the head (*jumping jacks*), and for very fast movements (*jumping jacks*).

E. Limitations

Regarding our study, several limitations have to be kept in mind. One limitation of our evaluation of K4ABT is that we compared it to our method because interference issues hindered a direct comparison to a marker-based system. Therefore, we need to treat these results with caution. Nonetheless, we only studied error metrics implying severe tracking failures, which we additionally confirmed by visual examination. Therefore, while we are confident that our results display the actual capabilities of K4ABT, another valuable approach would be to compare K4ABT with commercial markerless multi-camera systems like Theia markerless [40] or The Captury [44].

Including only healthy participants can be considered a limitation since the method will be applied to children with motor disorders. Nevertheless, to evaluate the tracking performance, all participants should be able to execute the movement tasks, including the challenging high-velocity tasks.

In this study, we evaluated the accuracy of 3D joint positions. We refrained from considering angles because differences in skeleton definitions have been reported to constitute a significant source of error for angle comparison [9], which would hinder an appropriate validation. While differences in skeleton definitions could also affect joint positions, we tried to reduce this as much as possible by applying the transformation from one skeleton to another for each joint separately instead of using one transformation for all joints at once. However, for clinical use, we will validate clinically more relevant parameters, such as joint angles, in a subsequent step. An advantage of our method is that it outputs the entire body surface, allowing a more detailed analysis of hand and foot contacts or distances to object surfaces.

Our method is based on the SMPL-H body model, which has the inherent drawback that the underlying skeleton is not anatomically correct. The joints of the skeleton serve as centers of rotation for body parts so that a realistic body surface is maintained when parts are rotated. Recent work introduced a first step towards integrating an anatomical skeleton inside a similar parametric body model [45].

VI. Conclusion

In this study, we evaluated the accuracy and concurrent validity of a custom method for markerless 3D full-body tracking of children and young adults from RGB-D sequences with respect to a marker-based Vicon system. We recorded 23 children and young adults performing five movement tasks. Our method's tracking results closely agree with those of the gold standard system: an average MPJPE over all joints of 11.7 mm, a percentage of correct keypoints of 98.4% for an error threshold of 50 mm, and very high ($r > 0.95$) to nearly perfect ($r > 0.99$) correlations in all dimensions if some motion or a lot of motion was present in a joint, respectively. We observed tracking failures (errors > 100 mm) in rare cases when segmentation failed due to motion blur or a full limb was completely hidden behind the body. The publicly available K4ABT tracking method showed overall good accuracy but, at the same time, recurring tracking errors. Therefore, we recommend restricting the clinical application of K4ABT to movements that can be easily tracked and verifying the correctness of kinematics.

We conclude that our method constitutes a new tool for accurate, portable, easy-to-use, markerless 3D motion tracking for children and adults that can enable motion analysis outside laboratories, e.g., at the patient's home.

We aim to integrate motion analysis based on our method into routine clinical examinations. We are currently conducting studies in which we complement standardized clinical assessments with our method to refine the quantification of motor function of children with neuromotor impairments. Consecutively, we will collect reference data on typically

developing children to develop automated methods for the objective, quantitative clinical assessment of motor function.

ACKNOWLEDGMENT

The authors would like to thank the participants and their parents and thank the therapists from their gait laboratory for their support. Nikolas Hesse provides consulting services to Meshcapade GmbH, but this research was performed solely with Swiss Children's Rehab. The other authors report no conflict of interest.

REFERENCES

[1] P. Eichelberger et al., "Analysis of accuracy in optical motion capture—A protocol for laboratory setup evaluation," *J. Biomech.*, vol. 49, pp. 2085–2088, Jul. 2016.
[2] S. Armand, G. Decoulon, and A. Bonnefoy-Mazure, "Gait analysis in children with cerebral palsy," *EFORT Open Rev.*, vol. 1, no. 12, pp. 448–460, 2016.
[3] H. Haberfehlner et al., "Instrumented assessment of motor function in dyskinetic cerebral palsy: A systematic review," *J. Neuroeng. Rehabil.*, vol. 17, p. 39, Mar. 2020.
[4] L. Mailleux et al., "Clinical assessment and three-dimensional movement analysis: An integrated approach for upper limb evaluation in children with unilateral cerebral palsy," *PLoS ONE*, vol. 12, pp. 1–24, Jul. 2017.
[5] M. C. M. Klotz et al., "Motion analysis of the upper extremity in children with unilateral cerebral palsy—An assessment of six daily tasks," *Res. Develop. Disabilities*, vol. 35, pp. 2950–2957, Nov. 2014.
[6] L. Heyrman et al., "Altered trunk movements during gait in children with spastic diplegia: Compensatory or underlying trunk control deficit?" *Res. Develop. Disabilities*, vol. 35, pp. 2044–2052, Sep. 2014.
[7] M. D. C. Vilas-Boas et al., "Validation of a single RGB-D camera for gait assessment of polyneuropathy patients," *Sensors*, vol. 19, no. 22, p. 4929, 2019.
[8] J. A. Albert, V. Owolabi, A. Gebel, C. M. Brahms, U. Granacher, and B. Arnrich, "Evaluation of the pose tracking performance of the Azure Kinect and Kinect v2 for gait analysis in comparison with a gold standard: A pilot study," *Sensors*, vol. 20, no. 18, p. 5104, 2020.
[9] L.-F. Yeung, Z. Yang, K. C.-C. Cheng, D. Du, and R. K.-Y. Tong, "Effects of camera viewing angles on tracking kinematic gait patterns using Azure Kinect, Kinect v2 and orbbec astra pro v2," *Gait Posture*, vol. 87, pp. 19–26, Jun. 2021.
[10] S. Springer and G. Y. Seligmann, "Validity of the Kinect for gait assessment: A focused review," *Sensors*, vol. 16, no. 2, p. 194, 2016.
[11] A. Pfister, A. M. West, S. Bronner, and J. A. Noah, "Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis," *J. Med. Eng. Technol.*, vol. 38, no. 5, pp. 274–280, 2014.
[12] Y. Ma, K. Mithraratne, N. Wilson, Y. Zhang, and X. Wang, "Kinect v2-based gait analysis for children with cerebral palsy: Validity and reliability of spatial margin of stability and spatiotemporal variables," *Sensors*, vol. 21, no. 6, p. 2104, 2021.
[13] B. F. Mentiplay et al., "Gait assessment using the Microsoft Xbox one Kinect: Concurrent validity and inter-day reliability of spatiotemporal and kinematic variables," *J. Biomech.*, vol. 48, pp. 2166–2170, Jul. 2015.
[14] K. Otte et al., "Accuracy and reliability of the Kinect version 2 for clinical measurement of motor function," *PLoS ONE*, vol. 11, pp. 1–17, Nov. 2016.
[15] A. Napoli, S. Glass, C. Ward, C. Tucker, and I. Obeid, "Performance analysis of a generalized motion capture system using Microsoft Kinect 2.0," *Biomed. Signal Process. Control*, vol. 38, pp. 265–280, Sep. 2017.
[16] R. A. Clark, B. F. Mentiplay, E. Hough, and Y. H. Pua, "Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives," *Gait Posture*, vol. 68, pp. 193–200, Feb. 2019.
[17] M. Antico et al., "Postural control assessment via Microsoft Azure Kinect DK: An evaluation study," *Comput. Methods Programs Biomed.*, vol. 209, Sep. 2021, Art. no. 106324.
[18] G. Sciortino, G. M. Farinella, S. Battiato, M. Leo, and C. Distante, "On the estimation of children's poses," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 410–421.
[19] J. Shotton et al., "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.
[20] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular RGB-D sequences," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2300–2308.
[21] G. Moon, J. Y. Chang, and K. M. Lee, "V2v-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5079–5088.
[22] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. G. Hofmann, and A. S. Schroeder, "Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2540–2551, Oct. 2020.
[23] L. Heyrman et al., "A clinical tool to measure trunk control in children with cerebral palsy: The trunk control measurement scale," *Res. Develop. Disabilities*, vol. 32, pp. 2624–2635, Nov./Dec. 2011.
[24] D. J. Russell, P. L. Rosenbaum, D. T. Cadman, C. Gowland, S. Hardy, and S. Jarvis, "The gross motor function measure: A means to evaluate the effects of physical therapy," *Develop. Med. Child Neurology*, vol. 31, pp. 341–352, Jun. 1989.
[25] F. Leboeuf, J. Reay, R. Jones, and M. Sangeux, "The effect on conventional gait model kinematics and kinetics of hip joint centre equations in adult healthy gait," *J. Biomech.*, vol. 87, pp. 167–171, Apr. 2019.
[26] Microsoft. (2021). *Azure Kinect Sensor SDK*. [Online]. Available: https://github.com/microsoft/Azure-Kinect-Sensor-SDK
[27] C. S. Bamji et al., "IMpixel 65 nm BSI 320 MHz demodulated TOF Image sensor with 3μm global shutter pixels and analog binning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 94–96.
[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[29] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Trans. Graph.*, vol. 36, no. 6, p. 245, 2017.
[30] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10975–10985.
[31] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13468–13478.
[32] *Open Source Tool for Making 3D Characters*, MakeHuman, 2021. [Online]. Available: http://www.makehumancommunity.org/
[33] S. Geman and D. E. McClure, "Statistical methods for tomographic image reconstruction," in *Proc. 46th Session Int. Stat. Inst., Bull. (ISI)*, 1987, pp. 5–21.
[34] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Computer Vision–(ECCV)*. Cham, Switzerland: Springer, 2016.
[35] J. Nocedal and S. J. Wright, "Nonlinear equations," in *Numerical Optimization*, New York, NY, USA: Springer, 2006, pp. 270–302.
[36] N. Ravi et al., "Accelerating 3D deep learning with PyTorch3D," 2020, *arXiv:2007.08501*.
[37] Microsoft. (2021). *Azure Kinect Body Tracking OfflineProcessor Sample*. [Online]. Available: https://github.com/microsoft/Azure-Kinect-Samples/tree/master/body-tracking-samples/offline_processor
[38] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 698–700, Sep. 1987.
[39] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[40] (2022). *Theia Markerless*. [Online]. Available: https://theiamarkerless.ca
[41] R. M. Kanko et al., "Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system," *J. Biomech.*, vol. 122, Jun. 2021, Art. no. 110414.
[42] R. Bashirov et al., "Real-time RGBD-based extended body pose estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2807–2816.
[43] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas, "HuMoR: 3D human motion model for robust pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11488–11499.
[44] (2022). *The Captury*. [Online]. Available: https://captury.com/
[45] M. Keller, S. Zuffi, M. J. Black, and S. Pujades, "OSSO: Obtaining skeletal shape from outside," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20492–20501.