Learning an Infant Body Model from RGB-D Data for Accurate Full Body Motion Analysis

Nikolas Hesse[1]*, Sergi Pujades[2], Javier Romero[3], Michael J. Black[2], Christoph Bodensteiner[1], Michael Arens[1], Ulrich G. Hofmann[4], Uta Tacke[5], Mijna Hadders-Algra[6], Raphael Weinberger[7], Wolfgang Müller-Felber[7], and A. Sebastian Schroeder[7]

[1]Fraunhofer Institute for Optronics, System Technologies and Image Exploitation, Ettlingen, Germany, [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany, [3]Amazon, Barcelona, Spain, [4]University Medical Center Freiburg, Faculty of Medicine, University of Freiburg, Germany, [5]University Children's Hospital Basel, Switzerland, [6]University of Groningen, University Medical Center Groningen, Netherlands, [7]Ludwig Maximilian University, Hauner Children's Hospital, Munich, Germany

Abstract. Infant motion analysis enables early detection of neurodevelopmental disorders like cerebral palsy (CP). Diagnosis, however, is challenging, requiring expert human judgement. An automated solution would be beneficial but requires the accurate capture of 3D full-body movements. To that end, we develop a non-intrusive, low-cost, lightweight acquisition system that captures the shape and motion of infants. Going beyond work on modeling adult body shape, we learn a 3D Skinned Multi-Infant Linear body model (SMIL†) from noisy, low-quality, and incomplete RGB-D data. We demonstrate the capture of shape and motion with 37 infants in a clinical environment. Quantitative experiments show that SMIL faithfully represents the data and properly factorizes the shape and pose of the infants. With a case study based on general movement assessment (GMA), we demonstrate that SMIL captures enough information to allow medical assessment. SMIL provides a new tool and a step towards a fully automatic system for GMA.

Keywords: body models, data-driven, cerebral palsy, motion analysis, pose tracking, general movement assessment

1 Introduction

One of the most common neurodevelopmental disorders in children is *cerebral palsy* (CP), which is caused by abnormal development of, or damage to the brain. Symptoms vary, but often include spasticity, abnormal muscle tone or impaired motor skills. Early intervention seems to have a positive effect on cognitive and motor outcome [18], yet requires early diagnosis. Neurological examinations or

*nikolas.hesse@iosb.fraunhofer.de

†SMIL is publicly available for research purposes at `http://s.fhg.de/smil`

technical assessment of brain functions show a large variation in predicting developmental outcome [5], and reliable diagnoses are generally obtained between the age of one and two years [19]. Prechtl discovered that the quality of spontaneous movements, in particular of the *general movements* (GMs), at the corrected age of 2-4 months accurately reflects the state of the infant's nervous system [15]. As of today, the *general movement assessment* (GMA) method achieves the highest reliability for the diagnosis and prediction of CP at such an early age [11]. Trained experts, usually physicians, analyze video recordings of infants and rate the GM quality, ranging from *normal optimal* to *definitely abnormal* in a modified version of Prechtl's GMA [5]. Infants with abnormal movement quality have very high risk of developing CP or minor neurological dysfunction [5]. Despite being the most accurate clinical tool for early diagnosis, GMA requires a trained expert and suffers from human variability. These experts need regular practice and re-calibration to assure adequate ratings. This motivates the need for automated analysis. To allow GMA automation, a practical system must first demonstrate that it is capable of capturing the relevant information needed for GMA. Moreover, to allow its widespread use, the solution needs to be seamlessly integrated into the clinical routine. Ideally it should be low-cost, easy-to-setup, and easy-to-use, producing minimal overhead to the standard examination protocol, and not affect the behavior of the infants.

We present the first work on 3D shape and 3D pose estimation of infants, as well as the first work on learning a statistical 3D body model from low-quality, incomplete RGB-D data of freely moving humans. We contribute (i) a new statistical *Skinned Multi-Infant Linear* model (SMIL), learned from 37 RGB-D low-quality sequences of freely moving infants, and (ii) a method to register the SMIL model to the RGB-D sequences, capable of handling severe occlusions and fast movements. Quantitative experiments show how SMIL properly factorizes the pose and the shape of the infants, and allows the captured data to be accurately represented in a low-dimensional space. With a case-study involving a high-risk former preterm study population, we demonstrate that the amount of motion detail captured by SMIL is sufficient to enable accurate GMA ratings. Thus, SMIL provides a fundamental tool that can form a component in a fully automatic system for the assessment of GMs. We make SMIL available to the community for research purposes.

We review related work in the fields of *medical analysis of infant motion* and *statistical body modeling*.

An overview of existing approaches for automating and objectifying the task of GMA is presented in [11]. For automated analysis, accurately capturing the motions of freely moving infants is key and has been approached in different ways. *Intrusive* systems rely on markers captured by camera systems [12], or on sensors attached to the infant's limbs, like electro-magnetical sensors [8] or accelerometers [6]. These approaches are highly accurate, since measurement units are directly connected to the limbs. However, the sensors/markers affect the infant's behavior. In addition, the setup and calibration of such systems can be cumbersome, the hardware is often expensive and the acquisition pro-

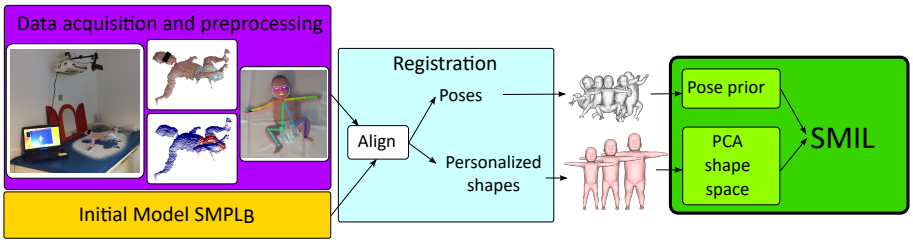

Fig. 1: Method overview. We acquire and preprocess RGB-D data. We create an initial infant model (SMPL_B) based on SMPL [9]. We register SMPL_B to the preprocessed data. We learn our new Skinned Multi-Infant Linear model (SMIL) consisting of a new shape space, and a new pose prior from the registrations.

tocol requires time consuming human intervention. *Non-intrusive* systems rely on simple, low-cost video or depth cameras, which facilitates usage in a broad clinical environment. From raw RGB videos, different body parts are tracked using optical flow [19] or weakly supervised motion segmentation techniques [16]. RGB-D sensors allow capturing motion in all three dimensions, e.g. by estimating joint positions based on a random ferns body part classifier [7]. Most similar to our work, the authors in [13] fit a body model consisting of simplistic shapes to RGB-D data and compare their method to sparse manually annotated landmarks. Differently to [13], we (i) learn a realistic infant body model from data, (ii) resolve rotational ambiguities by capturing full body shape and pose instead of 3D joint positions, and (iii) evaluate our model with surface distances, accounting for both pose and shape accuracy.

Statistical body models aim to describe the surface of humans or animals in a low-dimensional space. These models rely on sparse [1] or dense [9] surface data captured from cooperative, easy-to-instruct subjects or 3D toy models [21]. Infants present a major challenge in terms of data acquisition as they are not cooperative and cannot be instructed. Unlike previous work on human body models, we are not aware of a repository of high quality scans of infants, and thus, learn a 3D body model from RGB-D sequences of freely moving humans.

2 Learning the Infant Body Model

We create an initial infant model, $SMPL_\text{B}$, by adapting SMPL [9], and register it to the preprocessed data. Then, we learn our *Skinned Multi-Infant Linear* model (SMIL) from these registrations. The method overview is illustrated in Fig. 1. Manual intervention is only required in adjusting the pose priors (once for SMPL_B, once for SMIL), initial template creation (once for SMPL_B), and defining the number of clothing parts for each sequence (preprocessing).

Data Acquisition. We record freely moving infants for 3 to 5 minutes on the examination table without external stimulation, using a Microsoft Kinect V1 RGB-D camera. Ethics approval was obtained from Ludwig Maximilian Uni-

versity Munich (LMU) and all parents gave written informed consent for participation in this study.

Preprocessing. In the preprocessing step, we (i) transform depth images to 3D point clouds using the camera calibration, (ii) filter all table points not belonging to the infant by fitting a plane to the examination table, (iii) segment the infant point cloud into skin, diaper and onesie by adapting the segmentation method described in [14]. Finally, we (iv) extract landmarks from the RGB images, which provides us with 2D pose [4], hand locations [17] and facial landmarks [20], with their respective confidence estimates.

Initial Model. Learning an infant shape space is a chicken-and-egg problem: a model is needed to register the data, and registrations are needed to learn a model. We manually create our initial model SMPL_B, based on SMPL [9], a statistical body model learned from thousands of adult 3D scans. Simply scaling the adult model to infant size does not provide satisfactory results, as body proportions severely differ. We (i) replace the SMPL mean shape with an infant body mesh created with MakeHuman [10], (ii) leave the SMPL shape space untouched, (iii) scale the pose blendshapes to infant size, and (iv) manually adjust the pose priors. Because pose priors were learned on standing adults and not lying infants, adjusting these manually is important to prevent the model from explaining shape deformations with pose parameters.

Registration. The SMPL_B registrations to the preprocessed 3D point clouds are computed by minimizing the energy

$$\text{E}(\beta, \theta) = \text{E}_\text{data} + \text{E}_\text{lm} + \text{E}_\text{sm} + \text{E}_\text{sc} + \text{E}_\text{table} + \text{E}_\beta + \text{E}_\theta, \tag{1}$$

where E_data measures the scan to registration mesh distance, E_lm penalizes the distance between estimated and registration landmarks projected to 2D as in [3], E_sm enforces temporal pose smoothness and E_sc penalizes model self intersections as in [3]. E_table integrates background information in order to keep the bottom side of the registration body close to, but not inside the table. E_β and E_θ are the shape and pose prior, that enforce the shape parameters to be close to the mean, and help to prevent unnatural poses, respectively.

Initialization. Since the optimization problem is highly non-convex, the success of the registration depends on a good initialization. In contrast to adults, infants are incapable of striking poses on demand. Thus, relying on a predefined initial pose is unpractical. We overcome this by proposing a novel automatic method to select an initialization frame. We assume that a body segment is most visible if it has maximum 2D length over the sequence, since perspective projection decreases 2D body segment length. We choose the initialization frame as $f_\text{init} = \text{argmax}_f \sum_{s \in \text{S}} \text{len}(s, f) * c(s, f)$, where S is the set of segments, $\text{len}(s, f)$ is the 2D length of the segment s at frame f, and $c(s, f)$ is the estimated confidence of the joints belonging to s at frame f. For f_init we compute the initial registration by optimizing a simplified version of Eq. 1. It contains a 2D body pose landmark term similar to E_lm, a simplified data term, a strong prior on pose, and a shape regularizer. From f_init, we sequentially process the neighbouring frames (forward and backward in time), using as initialization the shape and pose results of the last processed frame.

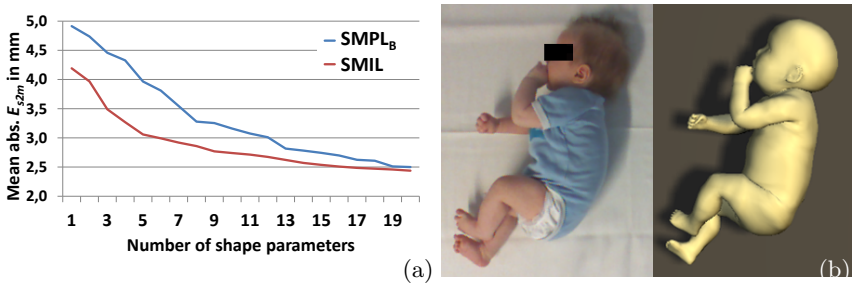

Fig. 2: (a): Average scan-to-mesh error E_{s2m} in mm w.r.t. the number of shape parameters for the two models aligned to all fusion scans. (b): example of RGB input image and the result of SMIL registered to the data.

Personalized Shape. For each sequence, we "unpose" the point clouds of a randomly selected subset of 1000 frames, similarly to [2]. The process of unposing changes the pose of the model into a normalized pose, which removes the variance related to body articulation. Because large parts of the infants' backs are never visible, we add model vertices that belong to faces oriented away from the camera, and call them *virtual points*. The union of the unposed scan points and the *virtual points* is the *fusion scan*. We register the model to the fusion scan by first optimizing only shape parameters and then optimizing for the free surface to best explain the fusion scan, by coupling the free surface to the first computed shape.

SMIL. To learn our *Skinned Multi-Linear Infant model*, we compute a new infant-specific *shape space* by doing weighted PCA on all 37 personalized shapes. We use low weights for points labeled as clothing and high weights for skin points, with smooth transitions in between, to avoid including diapers and clothing wrinkles in the shape space. We retain the first 20 shape components. In order to avoid repeated poses due to the lack of motion (sequences have between 4K and 10K frames), we randomly sample 1000 poses per sequence and learn the *pose prior* from 37K poses. As the learned prior does not penalize illegal poses (e.g. unnatural bending of knees) we manually add penalties to avoid them. Our SMIL model is composed of the shape space, the pose prior, and a base template, which is the mean of all personalized shapes.

3 Experiments

We evaluate SMIL quantitatively with respect to SMPL_B and perform a case-study on GMA ratings to demonstrate that SMIL captures enough information for medical assessment. Our dataset consists of 37 recordings of infants from a tertiary care high risk infants outpatient clinic, with an overall duration of over two hours. The infants' ages range from 9 to 18 weeks of corrected age (avg. of 14.6 weeks), their size range is 42 to 59 cm (avg. of 53.5 cm). We evaluate the SMIL model with a 9-fold cross-validation, using 33 sequences to train and 4 to

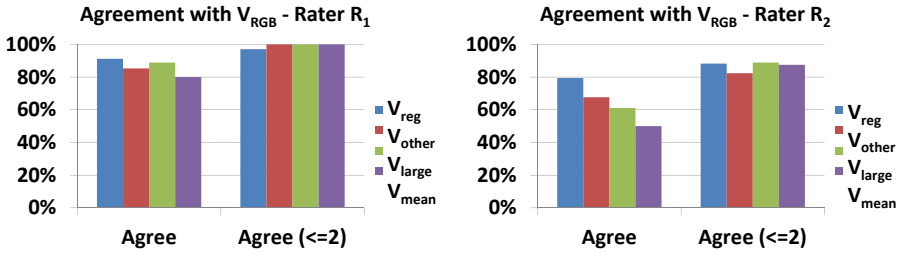

Fig. 3: Results of GMA case study. Percentage of ratings of synthetic sequences, generated using SMIL, that *agree* with the reference ratings R_1V_{rgb} (left) and R_2V_{rgb} (right), respectively. $V_{\{reg,other,large,mean\}}$ denotes different stimuli.

test. Different clothing styles (onesie, diaper, no clothing) are distributed across all sets. We evaluate the scan distance to the model mesh (E_{s2m}) by computing the Euclidean distance of each scan vertex to the mesh surface.

To evaluate the shape space, we register $SMPL_B$ and SMIL to the fusion clouds and evaluate E_{s2m} w.r.t. the number of shape parameters (Fig. 2a). SMIL is more accurate than $SMPL_B$ when using the same number of shape parameters. To evaluate how well the computed shapes and poses explain the input data we register $SMPL_B$ and SMIL to all sequences (200K frames) with the method described in Sec. 2 using 20 shape components. For $SMPL_B$, E_{s2m} is 2.67 mm (SD 0.22 mm), and for SMIL, E_{s2m} is slightly better: 2.51 mm (SD 0.21 mm). Fig. 2b shows a registration sample. Manual inspection of all sequences reveals 16 unnatural leg/foot rotations, lasting altogether 41 s (= 0.54% of total duration), and 18 failure cases (in 7 sequences), lasting altogether 49 s (= 0.66 % of total duration). The most common failure is "mixed up feet", i.e. feet aligned to the opposite side. Once, arm tracking is lost during side viewing, and one time a leg is severly twisted.

We conduct a case study on GMA to show that SMIL captures enough information to allow medical assessment. Three trained and certified GMA-experts perform GMA in different videos. We use five stimuli: i) the original RGB videos (denoted by V_{rgb}), and ii) the synthetic alignment videos (V_{reg}). For the next three stimuli we use the acquired poses of infants, but we animate a body with a different shape, namely iii) a randomly selected shape of another infant (V_{other}), iv) an extreme shape producing a very thick and large baby (V_{large}), and v) the mean shape (V_{mean}). We exclude three of the 37 sequences, as two are too short and one has non-nutritive sucking, making it non suitable for GMA. As the number of videos to rate is high (34*5), for iv) and v) we only use 50% of the sequences, resulting in 136 videos. For a finer evaluation, we augment standard GMA classes *definitely abnormal* (DA), *mildly abnormal* (MA), *normal suboptimal* (NS), and *normal optimal* (NO) [5] into a one to ten scale. Scores 1-3 correspond to DA, 4-5 to MA, 6-7 to NS, and 8-10 to NO. We consider two ratings with an absolute difference ≤ 1 to *agree*, and otherwise to *disagree*.

Rater R_1 is a long-time GMA teacher and has worked on GMA for over 25 years, R_2 has 15 years experience in GMA, and R_3 was certified one year ago, but lacks clinical routine in GMA. Average rating score (and standard deviation) for R1 is 4.7 (1.4), for R2 4.0 (1.9), and for R3 4.9 (2.3). The agreement on original RGB ratings V_{rgb} between R_3 and the more experienced raters is lower than 50%, while R_1 and R_2 agree on 65% of the ratings. This further stresses that GMA is challenging and its automation important. Due to the high rater variability we further focus on ratings of experienced raters R_1 and R_2. In Fig. 3, we present rating differences between synthetic and reference sequences. Each rater is compared to her own V_{rgb} ratings as a reference. R_1V_{reg} ratings *agree* on 91% of the reference ratings, whereas R_2 achieves an agreement rate of 79%. The agreement decreases more (R_2) or less (R_1) when the motions are presented with a different body shape. By extending the agreement threshold to ≤ 2, the percentages of all sequences become very similar. We intend to conduct further studies to elucidate the biases introduced by variation of shape.

4 Conclusions

In this paper, we contribute SMIL, a realistic, data-driven infant body model, learned from noisy, low-quality, incomplete RGB-D data, as well as a method to register SMIL to the data. Their combination allows the accurate capture of shape and 3D body motion of freely moving infants. Quantitative experiments showed that SMIL's metric accuracy is $\approx 2.5mm$. We demonstrated its clinical usability with a case study on general movement assessment. Our results illustrate the challenges of human GMA ratings - rater subjectivity and rater consistency - and reinforce the need for an automated system. Two experienced raters obtained 91% and 79% agreement between GMA ratings performed on original RGB videos and on synthetic videos generated using our method, indicating that SMIL captures enough motion detail for medical assessment. The introduction of shape variations led to a degradation of rating agreement.

Future work will study which non-motion related factors (body shape, texture, lighting) most affect the GMA ratings. Furthermore, we will target the automation of GMA by learning to infer ratings from the captured data. We are also investigating the usability of the system for quantification of disease progress and the impact of early therapy in infants with spinal muscular atrophy.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. ACM Trans. Graph. 24(3) (2005)
2. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In: IEEE International Conference on Computer Vision (ICCV) (2015)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: ECCV 2016. Lecture Notes in Computer Science, Springer (2016)

4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
5. Hadders-Algra, M.: General movements: a window for early identification of children at high risk for developmental disorders. The Journal of pediatrics 145(2) (2004)
6. 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) (2010)
7. Hesse, N., Stachowiak, G., Breuer, T., Arens, M.: Estimating body pose of infants in depth images using random ferns. In: IEEE International Conference on Computer Vision Workshops (ICCVW) (2015)
8. 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) (2008)
9. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. 34(6) (2015)
10. MakeHuman: Open source tool for making 3d characters. `www.makehuman.org`
11. 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)
12. 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) (2006)
13. Olsen, M.D., Herskind, A., Nielsen, J.B., Paulsen, R.R.: Model-based motion tracking of infants. In: ECCV Workshops. Springer (2014)
14. Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Trans. Graph. 36(4) (2017)
15. Prechtl, H.: Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. Early human development 23(3) (1990)
16. 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 (2015)
17. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
18. 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)
19. 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) (2012)
20. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
21. Zuffi, S., Kanazawa, A., Jacobs, D., Black, M.J.: 3D menagerie: Modeling the 3D shape and pose of animals. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)