IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 14, 2022, accepted March 9, 2022, date of publication March 21, 2022, date of current version April 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160749

RGBD Infant Head Reconstruction for Cranial Vault Asymmetry Estimation

SAMUEL ZEITVOGEL[1], **CHRISTIAN WERNET**[1], **JOHANNES WETZEL**[1], **ASTRID LAUBENHEIMER**[1], **AND KAI STOEVESANDT**[2]

[1]Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany
[2]VARILAG GmbH & Company KG DE, 76275 Ettlingen, Germany

Corresponding author: Samuel Zeitvogel (samuel.zeitvogel@h-ka.de)

This work was supported by the German Federal Ministry of Education and Research (BMBF) under Grant 13FH025IX6.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT We present an RGBD infant head reconstruction method with a mobile phone depth sensor on a novel dataset. Acquiring 3D models from infants enables many important medical tasks such as automatic cranial asymmetry classification for plagiocephaly therapy progress estimation. Existing methods for 3D infant head reconstruction employ synchronized multi-view configurations or hand-held laser scanning methods making their widespread employment difficult. In contrast, RGBD reconstruction methods either rely on static scenes failing on this task due to rapid infant head movements or employ dynamic methods lacking the high fidelity surface reconstructions required for accurate cranial measurements. We propose a domain-specific 3D reconstruction method augmenting static RGBD methods focusing on the rigid parts of the head and exploiting scene knowledge about the data acquisition methodology. We evaluate our approach using provided ground truth anthropometric measurements of the biparietal diameter and report competitive accuracy.

INDEX TERMS Anthropometry, RGBD reconstruction, segmentation, structure from motion.

I. INTRODUCTION

Starting in the early 90s, it is recommended that infants sleep on their backs to mitigate the Sudden Infant Death Syndrome (SIDS) [1]. While keeping their infants on their backs successfully reduces the risk of SIDS [2], studies hint at an increase in an infant's head shape deformation (plagiocephaly), favored by a homogeneous sleeping posture [3], [4]. Plagiocephaly refers to the medical condition of the skull deformation caused by persistent external directional forces [5]. Studies indicate that 37 to 46 percent [6], [7] of the infants 7 to 12 weeks of age suffer from plagiocephaly. While plagiocephaly can be successfully treated, unaddressed plagiocephaly can lead to a multitude of risk factors for the infant [8]–[12].

Common treatments are repositioning strategies [13], physical therapy [13], and cranial remolding orthoses [14]. During treatment, the therapy progress (i.e. the state of the infant's skull deformation) is supervised by medical experts, using specialized measurement tools, such as skull calipers [15], laser scanners [16], or multi-view 3D camera setups [15]. The listed tools require additional hardware, a trained operator and are potentially uncomfortable for the infant. For these reasons, a ubiquitous measurement method for therapy progress tracking that could even be used by untrained personnel such as the infant's parents would be desirable.

In this work, a 3D infant head reconstruction method and cranial asymmetry measurement method using a mobile phone is proposed. The advent of 3D sensor-equipped mobile phones such as the iPhone TrueDepth camera [17] enables widespread 3D reconstruction and measurement applications such as room measurements for interior design preview and object reconstruction.

The task of 3D infant head reconstruction comes with several challenges: Due to the sensitive data and the recent availability of phones equipped with depth sensors, publicly available data sets are hard to come by. Next, reconstructing

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

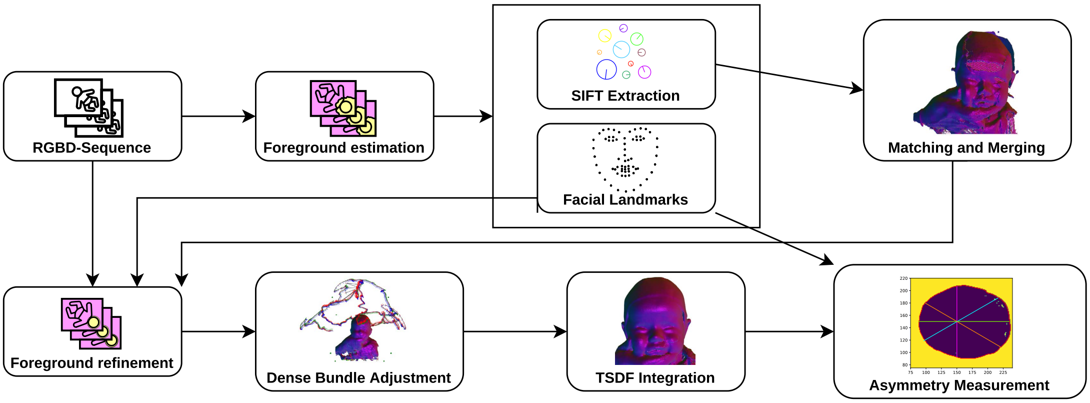
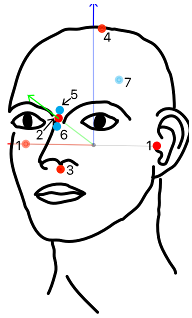


FIGURE 1. Overview of the proposed pipeline. The line segments depicted in the cross-section of node "Asymmetry Measurement" are color-coded as follows: AP (green), BP (violet), DIAG1 (blue), and DIAG2 (orange). The circumference is colored red.



1	left and right tragion
2	nasion
3	subnasale
4	vertex
5	glabella
6	sellion
7	opisthocranion
gray	origin (center of tragions)
green	y-axis through origin and 2
blue	z-axis in plane (origin,2,3)
red	resulting x-axis

FIGURE 2. Key landmarks and reference plane (x-y plane). The red points denote keypoints that are used to compute the reference plane following [19], [20]. Blue points denote keypoints that are used by other methods [15].

an infant's head with an RGBD sequence is hard due to the uncooperative behavior of the subject. Single-sensor 3D reconstruction methods can be simplified considerably when the reconstruction target is rigid and stationary [18]. Both assumptions do not hold when trying to reconstruct the head of an infant due to non-rigidity in the face area arising from facial expressions and sudden rapid, uncontrolled body movement of the infant. Additionally, the resulting 3D models often suffer from 3D artifacts making automatic measurement extraction challenging. In contrast, cranial vault measurements are performed on the rigid part of the human head which could be significantly easier to reconstruct.

Our contribution is twofold. First, we record 8 RGBD sequences, spanning a variety of infant heads with different head asymmetries. For each subject, cranial measurements were recorded manually by experts using specialized calipers. Second, we extend state-of-the-art RGBD reconstruction methods to generate 3D shape models of the head, without the aforementioned inconvenience and effort of existing methods for supervision.

II. RELATED WORK

A. MEASUREMENT TECHNOLOGIES

Direct methods for cranial vault asymmetry measurements are conducted by an expert using a caliper [15]. In contrast, a multitude of digital methods have been proposed. Plank *et al.* [20] use a non-invasive laser shape digitizer to extract a 3D shape model. Scan acquisition was conducted by putting a stockinet over the head of the infant. Meyer-Marcotty *et al.* [19] employ a non-invasive 3D scanning solution consisting of five synchronized depth sensors with an infant situated in the center. Jelinek *et al.* [21] use magnetic resonance imaging (MRI) and extract the relevant cross-sections to determine the cranial vault asymmetry. Skolnick *et al.* [15] compare direct caliper measurements with digital photogrammetric measurements.

Digital methods define a coordinate system constructed from key anatomical landmarks (depicted in Fig. 2). The x-y plane is denoted as the reference plane. A measurement plane is constructed parallel to the reference plane by offsetting the reference plane in the z-direction. Different methods to determine the offset are used: Meyer-Marcotty *et al.* [19] choose the offset in such a way, that the head perimeter is maximized. Plank *et al.* [20] determine the vertex height with respect to the reference plane and set the measurement plane offset to 0.3 times the vertex offset. A pair of diagonals within the measurement plane is constructed (see Fig. 1) and their lengths are used to determine the cranial vault asymmetry.

B. IMAGING SENSORS

In the family of digital 3D measurement methods, camera-based sensors are low-cost and ubiquitous. In this context, three sensor modalities are discussed in the context of 3D shape reconstruction: Synchronized multi-sensor networks, single moving RGB sensors, and single moving RGBD sensors.

Synchronized multi-sensor reconstruction methods are among the most accurate solutions. Because all data is acquired at the same time, even moving and deforming objects can be reconstructed on a frame-by-frame basis.

A single moving camera is among the most accessible solutions because every modern smartphone is equipped with a camera. Structure-from-motion (SfM) algorithms can deliver 3D reconstructions in this sensor setup but require sufficiently textured objects to estimate the camera trajectory reliably. While this is not an issue in the facial area, the missing texture near the scalp is a challenging area to reconstruct with RGB data only. To overcome these challenges, Barbero-García *et al.* [22] employ a textured cap using an off-the-shelf tool for a photogrammetry pipeline.

Active RGBD sensors can address this issue by additionally providing range information for most pixels by e.g. projecting infrared dot patterns on the target [23]. Additional challenges stemming from varying facial expressions and sporadic head movements have to be addressed by the employed 3D reconstruction method.

C. RGBD 3D SHAPE RECONSTRUCTION

A large body of work exists for RGBD 3D shape reconstruction. Reconstruction methods can be divided into static scene reconstruction and dynamic scene reconstruction. In the seminal work by Newcombe *et al.* [24] the KinectFusion algorithm is proposed enabling real-time surface reconstruction without using any color information. The sensor pose of an incoming frame is obtained via the iterative closest point (ICP) method using all depth data from earlier frames. Registered frames are integrated into a global scene model encoded with a truncated signed distance function (TSDF) [24] by converting each depth frame into a projective TSDF and aggregated with weighted running averaging.

A plethora of extensions for KinectFusion have been developed [18]. Dai *et al.* [25] employ sparse SIFT keypoints [26] in conjunction with dense correspondences in a global bundle adjustment objective to improve global model consistency. An involved correspondence filtering and dense verification step is performed. The current scene model is encoded using a TSDF and a feedback loop back into the correspondence estimation step is set up resulting in high-quality surfaces with robust camera tracking.

In contrast, dynamic 3D scene reconstruction methods model the surface deformation of objects in the scene [27]–[29]. Zollhöfer *et al.* [18] review the literature on the topic and classify dynamic 3D scene reconstruction as inherently ill-posed, and identify the increased number of unknowns, fast motions and real-time performance. Additional prior knowledge of the object class can improve the dynamic shape reconstruction process. Hesse *et al.* [30], [31] use a modified version of the statistical human shape model SMPL [32] to recover a watertight mesh from an RGBD sequence of an infant lying on an examination table and undergoing rapid body movement. The baby is segmented by exploiting the controlled recording setup and fitting a plane to the table.

Given that the skull is rigid and the additional difficulties that entail the employment of dynamic 3D scene reconstruction methods, we instead leverage static 3D scene reconstruction methods segmenting the rigid parts of the baby head and estimating the camera trajectory not with respect to the static parts of the scene but the rigid moving head shape. We drop the real-time performance aspect to simplify the algorithm design process leaving real-time capabilities to future work. Our approach draws from [25], injecting custom head segmentation functionality into the method to deal with the movement in the scene as well as the dynamic background. We adjust relevant hyperparameter settings to adapt a method that was used for room-scale reconstruction to close up head reconstruction.

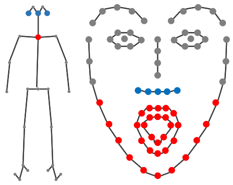


FIGURE 3. Mask refinement via OpenPose landmarks: OpenPose body landmarks (left) and facial landmarks (right) are depicted by dots of different colors. Red dots populate the set L^-, blue dots populate the set L^+ and gray dots are not used.

III. DATA ACQUISITION

We create a dataset of infants diagnosed with plagiocephaly. The whole dataset contains RGBD sequences from 8 infants. Each infant's parent gave written consent for use in a scientific study. For data capture, an iPhone 12 Pro running Record3D[1] to access the TrueDepth camera is used, also providing the camera matrix $K \in \mathbb{R}^{3\times3}$. Each RGBD sequence consists of 1248 frames on average at 30 frames per second. The infant was held by their parent and the operator was recording the video, slowly moving the camera around the infant while trying to keep a distance of roughly 30 cm to the target. Different head stabilization grips are present in the dataset with the parent stabilizing the head near the chin or the back of the neck. Due to the infant's movement and to avoid injuries, the supporting grip is not always rigid throughout a recording session. Infants with significant hair growth are wearing a skin-tight cap. The footage contains spurious head movements, body motions, and changing facial expressions making 3D reconstruction challenging. Sometimes the infant moves the head too close to the sensor resulting in several frames without valid depth data.

IV. METHOD

A. SYSTEM OVERVIEW

The stages of our proposed reconstruction pipeline are depicted in Fig 1. First, for all frames, initial foreground masks are estimated. Then, SIFT keypoints and facial landmarks are extracted in the foreground region. Then, SIFT keypoints are extracted and matched pair-wise. Correspondences are filtered with a random sample consensus (RANSAC) [33] approach. Global camera poses are estimated incrementally employing a correspondence count-based greedy merge strategy. The preliminary camera trajectory in conjunction with the facial landmarks is used to refine the foreground masks retaining only 3D information on the rigid parts of the head. The camera poses are finalized by minimizing a global energy function containing pair-wise dense correspondences. In the final step, all depth images are integrated into a TSDF and key measures classifying the head asymmetry are extracted.

[1] https://record3d.app/

B. KEYPOINT AND FOREGROUND EXTRACTION

In an RGBD sequence of $N = 1500$ consecutive frames the i-th RGB image is denoted by I^i and the i-th depth map is denoted by D^i for all $1 \leq i \leq N$. A pixel coordinate with present depth value is denoted by $(u, v) \in \Omega^i$, where Ω^i denotes the set of all pixels of frame i with valid depth values. We apply OpenPose [34] head keypoint extraction [35] on each RGB image I^i resulting in keypoints L^i for each frame filtering by detection score and discarding all keypoints with invalid depth values, and depth values above 50 cm. This filtering is done to discard keypoints stemming from other persons in the scene. Additionally, an initial foreground mask is estimated by thresholding the depth map in an adaptive fashion for each frame:

First, for a depth frame D^i the closest pixel of the infant's head is estimated by first identifying all depth values that are close to the sensor (< 50 cm). Then for the remaining valid depth values the depth value that is the 1% percentile is estimated to be among the closest pixels on the head with respect to the camera. This closest depth is denoted by $\bar{d}_i \in \mathbb{R}$. The initial mask is then computed by thresholding D^i by $\bar{d}_i + 0.1$m. The resulting binary foreground masks are denoted by M^i.

C. SPARSE CORRESPONDENCES

In the next step, SIFT features are extracted from each image I^i within the segmented region specified by M^i. Due to the rapid head movement observed in many frames, it is important to only extract SIFT features on the foreground because the camera movement with respect to the static objects is different from the camera movement with respect to the moving infant head. We deviate from the SIFT keypoint detection parameters formulated in [26] with a contrast threshold of 0.002, edge threshold of 20, and $\sigma_{\text{SIFT}} = 0.8$. This results in many SIFT keypoints also covering areas of lower contrast. This is typically not beneficial for keypoint matching, resulting in many incorrect keypoint correspondences. The constrained camera movement and object distance results in effective correspondence filtering opportunities discussed below.

After SIFT extraction, the relative camera pose to the head is estimated between pairs of frames by SIFT keypoint matching. A brute-force matching would result in $N(N-1)/2$ pairwise matching steps which is prohibitively computationally demanding. Instead, a sparse matching graph is constructed containing edges between temporally adjacent frames as well as temporally distant frames. The undirected graph consisting of N nodes $X = \{1, \cdots, N\}$ and $|C|$ edges is denoted by $C = C_{\text{close}} \cup C_{\text{far}}$ with

$$C_{\text{close}} = \{(i,j) \in X^2 |\ 1 \leq |j - i| \leq 5\} \tag{1}$$

and

$$C_{\text{far}} = \{(i,j) \in X^2 |\ i \neq j \wedge (i \mid 5) \wedge (j \mid 5)\}, \tag{2}$$

where $i \mid 5$ tests whether i is divisible by 5.

Given a pair (i, j) going to be matched, the corresponding SIFT keypoints are matched using nearest neighbor search resulting in $n_{i,j}$ matches. Given $n_{i,j} \geq 3$ linearly independent points, the 3D coordinates in the local camera coordinate systems of frames i and j are denoted by $\boldsymbol{p}_k^{i,j}, \boldsymbol{p}_k^{j,i} \in \mathbb{R}^3$ respectively for all $1 \leq k \leq n_{i,j}$. The relative camera poses are recovered by minimizing the reconstruction error

$$R_{i,j}, \boldsymbol{t}_{i,j} = \underset{R \in SO(3), \boldsymbol{t} \in \mathbb{R}^3}{\arg\min} \sum_{k=1}^{n_{i,j}} \left\| R\boldsymbol{p}_k^{i,j} + \boldsymbol{t} - \boldsymbol{p}_k^{j,i} \right\|_2^2, \tag{3}$$

resulting in the relative rigid transformations denoted by the rotation $R_{i,j}$ and translation $t_{i,j}$. The objective (3) can be minimized using the Kabsch algorithm [36]. The objective (3) is not robust to outliers and a RANSAC [33] inspired correspondence filtering method similar to [25] is incorporated exploiting the camera geometry and scene-knowledge.

D. ROBUST KEYPOINT MATCHING

The RANSAC correspondence matching algorithm is derived from [25]. From all correspondence candidates, a subset of three correspondences is drawn and an initial rigid transform is computed with the Kabsch algorithm. Then, a number of correspondence inliers that conform with the estimated transformation is identified. The whole process is repeated 1000 times and the transformation resulting in the most inliers is identified. Transformation candidates resulting in less than five inliers are rejected. All correspondences classified as outliers are discarded.

A correspondence is classified as an inlier if and only if the pair of points are within 5 mm after applying the estimated transformation. A transformation estimate with a rotation angle of more than 30° or a translation distance of more than 20 cm is rejected outright. Additionally, a good spread of the correspondences across the head is required to ensure a stable rigid transformation estimate. To this end, the covariance matrix of the centered source correspondences is computed and the condition number is calculated. A condition number larger than 100 is deemed unstable and the transformation is rejected. The same test is performed on the covariance matrix of the centered target correspondences. Next, the cross-covariance between the centered source points and centered target points is formed and the condition number is tested.

In the final validation step, a surface area test is performed for the source and target point clouds. We describe the procedure for the source point cloud. The source point cloud is projected to the plane spanned by the two eigenvectors corresponding to the two largest eigenvalues of the centered source point cloud covariance matrix. From the resulting 2D point cloud the convex hull is determined. If the area of the convex hull is below 16 cm^2, the transformation is rejected.

All successfully matched pairs are encoded by the graph with vertices $X' \subset X$ and edges $C' \subseteq C$. Using the graph defined by C' a greedy incremental merging strategy is devised. The subgraph formed by the largest connected

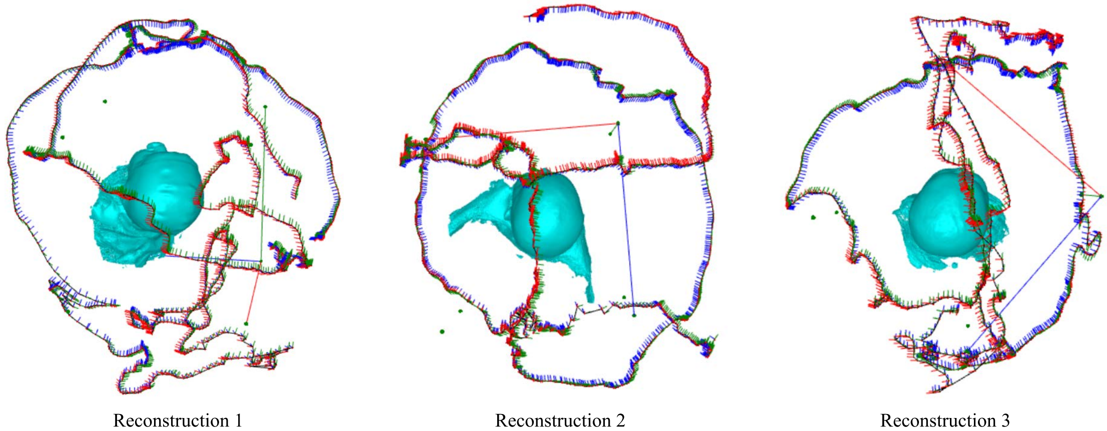


FIGURE 4. 3D reconstructions and camera trajectory relative to the reconstructed head of the infant.

Algorithm 1 Greedy Frame Merging Strategy

```
Require: Camera visibility graph ($X''$, $C''$)
$X_{\text{merge}} = [x_0]$
while $|X_{\text{merge}}| < |X''|$ do
    $x \leftarrow \arg\min_{i \in X'' \setminus X_{\text{merge}}} \sum_{j \in X_{\text{merge}}} n_{i,j}$
    Append $x$ to $X_{\text{merge}}$
end while
return Merging order $X_{\text{merge}}$
```

component is denoted by the vertices $X'' \subset X'$ and edges $C'' \subseteq C'$. First, the biggest connected component is extracted and the remaining frames $X \setminus X''$ are discarded. The merging order X_{merge} is determined following Alg. 1.

After determining the merging order, incremental camera pose estimation is performed. Here, we describe a single step of the merging approach that adds a single frame to the scene. Given a set of frame indices X'', the set of already merged frame indices $S \subseteq X''$, the camera pose estimates $R_j \in SO(3)$ and $\boldsymbol{t}_j \in \mathbb{R}^3$ for all $j \in S$, the next frame index i in the merging order with $i \notin S$ and

$$S_i = \{j \in S | (i, j) \in C''\}, \tag{4}$$

the rigid camera-to-world transformation specified for frame index i is denoted by $R_i \in SO(3)$ and $\boldsymbol{t}_i \in \mathbb{R}^3$ and computed via

$$R_i, \boldsymbol{t}_i = \underset{R \in SO(3), \boldsymbol{t} \in \mathbb{R}^3}{\arg\min} \sum_{j \in S_i} \sum_{k=1}^{n_{i,j}} \left|\left| R\boldsymbol{p}_k^{i,j} + \boldsymbol{t} - (R_j \boldsymbol{p}_k^{j,i} + \boldsymbol{t}_j) \right|\right|_2^2 \tag{5}$$

Analogous to (3), (5) can be minimized with the Kabsch algorithm.

E. FOREGROUND MASK REFINEMENT

After merging is finished, there exists a pose estimate for each frame index. In the next step, the foreground masks are refined exploiting gained knowledge of the scene through the merging process. A depiction of this process is supplied in Fig. 5. To be more precise, the camera trajectory, and the OpenPose keypoints are used to refine the masks considerably. To this end, the auxiliary camera positions $\boldsymbol{t}'_i$ are defined using trajectory center $\bar{\boldsymbol{t}}$ with

$$\boldsymbol{t}'_i = \boldsymbol{t}_i + 0.35(\bar{\boldsymbol{t}} - \boldsymbol{t}_i), \tag{6}$$

enveloping the head of the infant. The auxiliary camera positions, as well as the OpenPose 3D keypoints of the face are combined to form an auxiliary point cloud Q. Then the convex hull of Q is calculated. Each depth value of each image is transformed into the world coordinate system via R_i and t_i. Subsequently, each point is tested whether it lies within the convex hull. Pixels within the convex hull are marked as foreground pixels and the pixels outside of the convex hull are classified as background pixels.

Due to large temporal shape variability in the mouth area, we try to remove the corresponding pixels from the foreground mask. Pixels below the plane containing the subnasale, left tragion and right tragion are classified as background pixels. To construct the separation plane automatically, landmark point clouds $L^+, L^- \subset \mathbb{R}^3$ are constructed exploiting the predicted landmark positions in space. The set L^+ contains points that are above the separation plane and L^- contains points that are below the separation plane (see Fig. 3). The separation plane is estimated seeking the max-margin class separation plane. This is achieved by training a linear support vector machine (SVM) [37] with positive samples L^+ and negatives samples L^- respectively. This results in our finalized foreground masks $\hat{M}_i$ for all frames $1 \leq i \leq N$. The mask refinement process is depicted in Fig. 5.

F. DENSE CORRESPONDENCES

The initial camera pose estimates are computed using the sparse SIFT feature matches. The predicted camera trajectory might not be accurate enough for 3D reconstruction. For this reason, a dense correspondence term is introduced and minimized globally by optimizing all frame camera poses jointly. The sparse SIFT keypoints are discarded but the initial camera trajectory is used as an optimization starting configuration and refined to enable detailed 3D surface reconstruction.

The dense correspondence energy formulation is split into two parts, a pixel candidate selection followed by a sum over pairwise dense geometric image energy.

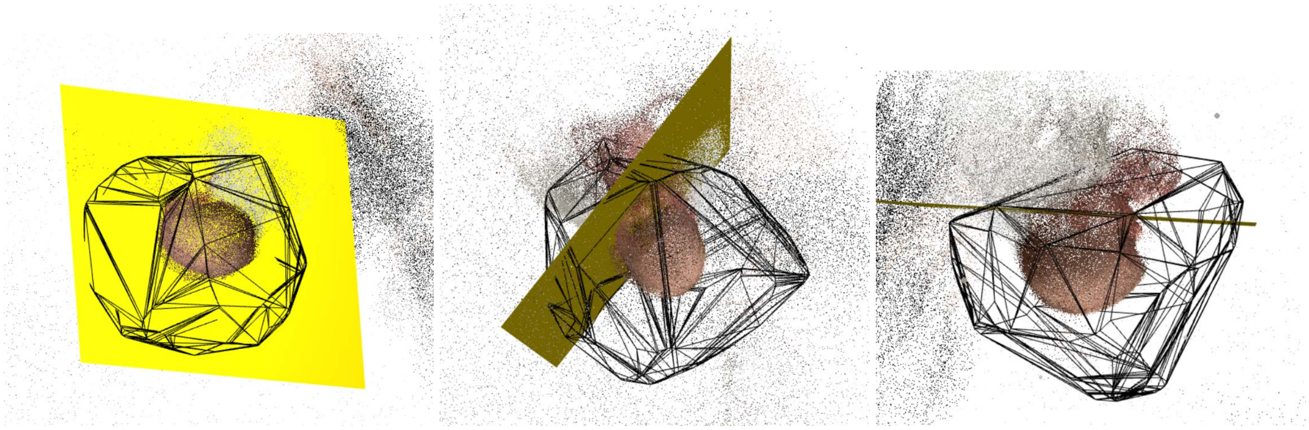


FIGURE 5. Geometric depiction of the mask refinement process. The estimated camera trajectory is shifted towards the center of all camera positions. Then the convex hull is computed (shown by black lines) and all points outside of the convex hull are marked as background pixels. For facial keypoints a separation plane (depicted in yellow) is estimated using a linear support vector machine. All points that fall below the separation plane are marked as background pixels.

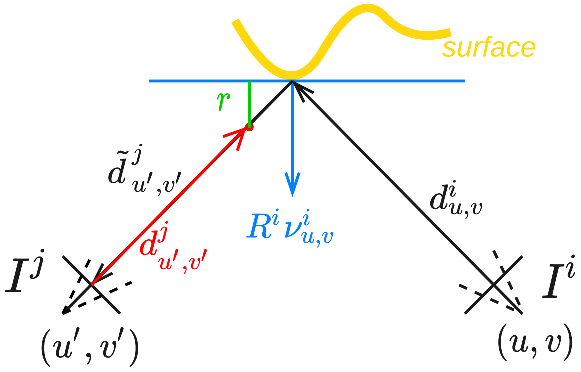


FIGURE 6. Dense pixel association.

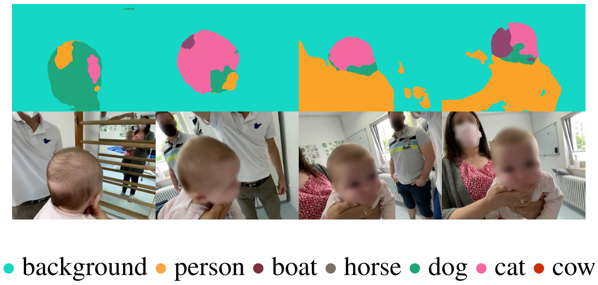


FIGURE 7. Segmentation results of the Pyramid Scene Parsing Network [38] on selected frames of the recorded dataset. The network seems to interpret the infant's hair as animal fur or background.

For the pixel candidate selection, a normal image is calculated from each depth map by estimating a plane from the 3 × 3 neighborhood of each pixel. Per image pair candidate pixel selection is computed by shooting rays from the source image into the scene and projecting them onto the target image. Depth disparity, color disparity, and normal deviation are thresholded resulting in a binary mask of valid dense pixel correspondences. The source-to-target mapping is for a pixel (u, v) in frame i to a pixel (u', v') in frame j is computed via

$$\tilde{d}^j_{u',v'}\begin{pmatrix}u'\\ v'\\ 1\end{pmatrix} = KR_j^{-1}\left(R_iK^{-1}d^i_{u,v}\begin{pmatrix}u\\ v\\ 1\end{pmatrix} + \boldsymbol{t}_i - \boldsymbol{t}_j\right), \quad (7)$$

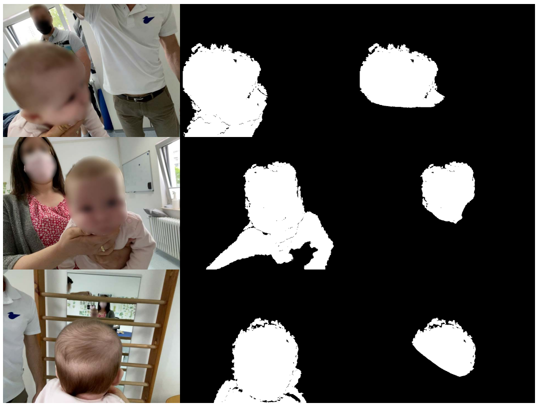


FIGURE 8. Segmentation results before and after refinement. From left to right: color image, initial foreground estimation, and refined foreground mask. Each row shows a different view of the same reconstruction target. The color image is only used for visualization purposes.

where $d^i_{u,v}$ denotes the depth value of pixel location (u, v) of the source frame i and $\tilde{d}^j_{u',v'}$ denote the predicted depth of the corresponding target frame pixel location (u', v'). Note, that u' and v' are functions of R_i, R_j, $\boldsymbol{t}_i$, and $\boldsymbol{t}_j$. Given the RGB values $\boldsymbol{I}^i_{u,v}, \boldsymbol{I}^j_{u',v'} \in [0, 1]^3$, normal vectors $\boldsymbol{\nu}^i_{u,v}, \boldsymbol{\nu}^j_{u',v'} \in \mathbb{R}^3$, and depth values $d^i_{u,v}, d^j_{u',v'} \in \mathbb{R}$ a valid candidate has to satisfy the following conditions [25]. We denote a pixel for a pair of frames (i, j) valid by $\nu^{i\rightarrow j}_{u,v} \in \{0, 1\}$ with

$$\begin{aligned}\nu^{i\rightarrow j}_{u,v} &= |\boldsymbol{I}^i_{u,v} - \boldsymbol{I}^j_{u',v'}|_1 \leq \lambda_{\text{color}} \;\wedge\\ (\boldsymbol{\nu}^i_{u,v})^T\boldsymbol{\nu}^j_{u',v'} &\geq \lambda_{\text{normal}} \;\wedge\\ |d^j_{u',v'} - \tilde{d}^j_{u',v'}| &\leq \lambda_{\text{depth}}. \end{aligned} \quad (8)$$

In case the coordinates (u', v') are not within the frame j or no valid depth $d^j_{u',v'}$ or normal $\nu^j_{u',v'}$ exists, the dense correspondence is also rejected and $\nu^{i\rightarrow j}_{u,v} = 0$. RGB values, normal vectors, and depth values are interpolated bi-linearly at the coordinates (u', v'). Depth discontinuities are filtered by λ_{normal}. The threshold parameters $\lambda_{\text{color}}, \lambda_{\text{normal}}, \lambda_{\text{depth}} \in \mathbb{R}^+$ are adaptively refined during the dense alignment procedure following a coarse-to-fine strategy (see Section IV-G).

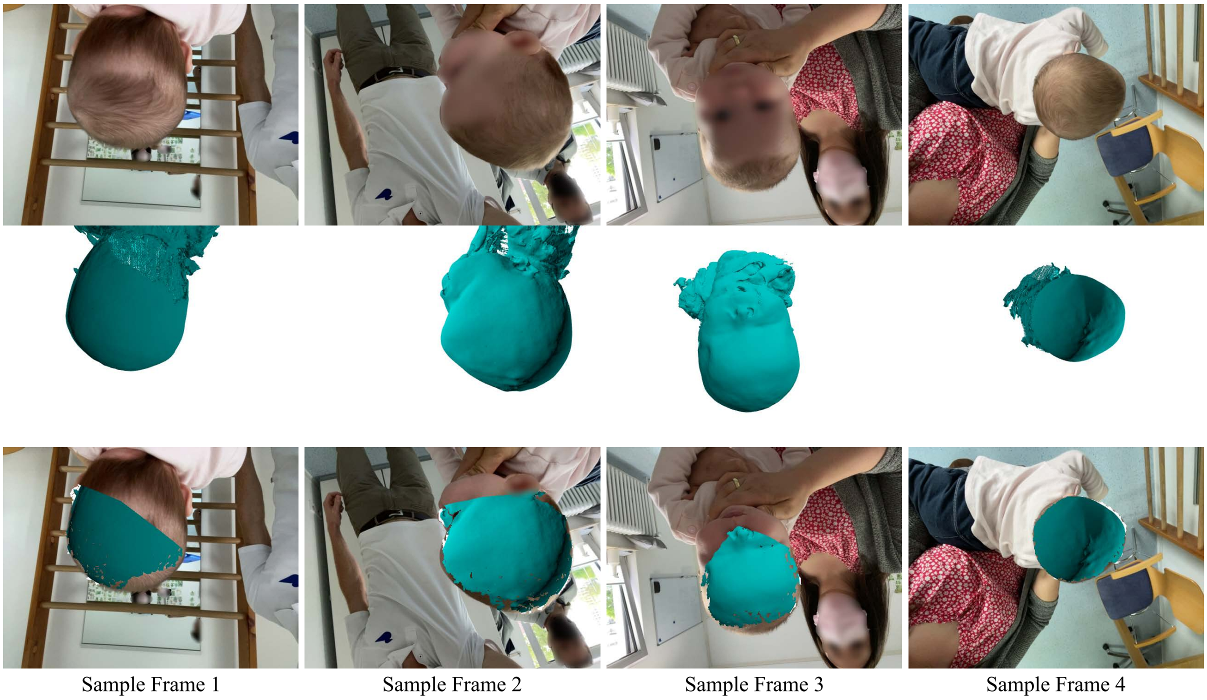


FIGURE 9. 3D reconstruction results for one sequence: The first row shows four RGB images of a recorded sequence. The second row shows the corresponding 3D reconstruction projected into the respective frame. The third row shows an overlay of the rendered view and the RGB image. Pixels from the RGB image are only replaced for the pixels specified by the refined foreground mask $\hat{M}_i$.

We denote the set of valid pixels for the frame pair (i, j) by $V^{i\to j}$ with

$$V^{i\to j}_{\text{unfiltered}} = \{(u, v) \in \Omega_i | v^{i\to j}_{u,v} = 1\}.$$

$$V^{i\to j} = \begin{cases} V^{i\to j}_{\text{unfiltered}} & \text{if } |V^{i\to j}_{\text{unfiltered}}| \geq 1000 \\ \emptyset & \text{else,} \end{cases} \tag{9}$$

discarding dense correspondences altogether, if less than 1000 dense correspondences are found. Equipped with a dense correspondence filtering strategy, a dense correspondence energy function E can be defined. A geometric visualization providing intuition is depicted in Fig. 6. Following the dense geometric error formulation in [25] and denoting $\theta = (R_1, , \cdots, R_N, \boldsymbol{t}_1, \cdots, \boldsymbol{t}_N)$, we write

$$E(\theta) = \sum_{(i,j)\in C} \sum_{(u,v)\in V^{i\to j}} r^{i\to j}_{u,v}(\theta)^2 \tag{10}$$

with

$$r^{i\to j}_{u,v}(\theta) = (\boldsymbol{v}^i_{u,v})^T \left(\boldsymbol{p}^{i\to j}_{u,v} - K^{-1} d^i_{u,v} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \right) \tag{11}$$

and

$$\boldsymbol{p}^{i\to j}_{u,v} = R_i^{-1} \left(R_j K^{-1} d^j_{u',v'} \begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} + \boldsymbol{t}_j - \boldsymbol{t}_i \right). \tag{12}$$

G. OPTIMIZATION

Following up on the introduced energy E, in this section, a numerical minimization scheme is devised to find adequate camera poses $(\hat{R}_i, \hat{\boldsymbol{t}}_i)$ for all frames $1 \leq i \leq N$. To fix the global transformation, we set the transformation of the first frame to $R_1 = I$ and $\boldsymbol{t}_1 = \boldsymbol{0}$ and seek the camera poses $\hat{\theta} = (\hat{R}_2, \cdots, \hat{R}_N, \hat{\boldsymbol{t}}_2, \cdots, \hat{\boldsymbol{t}}_N)$ minimizing E denoted by

$$\hat{\theta} = \underset{R_2,\cdots,R_N,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_N}{\arg\min} E(R_1, \cdots, R_N, \boldsymbol{t}_1, \cdots, \boldsymbol{t}_N). \tag{13}$$

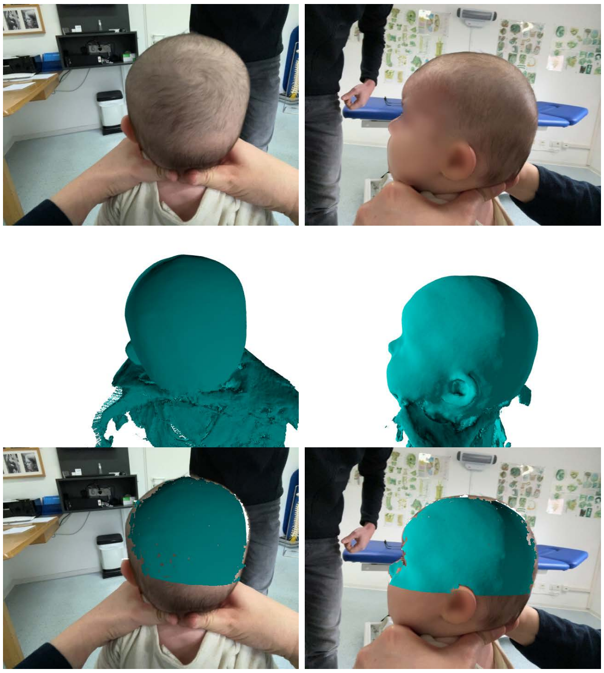
The structure of objective function E allows the employment of non-linear least squares solvers. Assuming an adequate initialization of the camera trajectory provided by the sparse correspondence pipeline, the Gauss-Newton method is used for objective minimization. The residuals and Gauss-Newton matrix are assembled in parallel on an NVIDIA GeForce RTX 3090 using the JAX library [39]. Rotations are optimized by linearization assuming small angles [24].

To avoid costs arising from misaligned frames in the objective we remove the variables R_i and $\boldsymbol{t}_i$ from the objective if the frame i is disconnected within the objective (10) and would prevent the Gauss-Newton matrix from factorizing successfully using sparse Cholesky decomposition.

Dense alignment is performed with a coarse-to-fine strategy, running the Gauss-Newton algorithm for 60 iterations. The optimization is performed in three stages running for 20 iterations each varying the threshold parameters λ_{color}, λ_{normal} and λ_{depth} (see Table 1). The current implementation optimizes all camera poses jointly without any real-time capable hierarchical optimization (see e.g. [25]) and runs for approximately 4 hours per sequence.

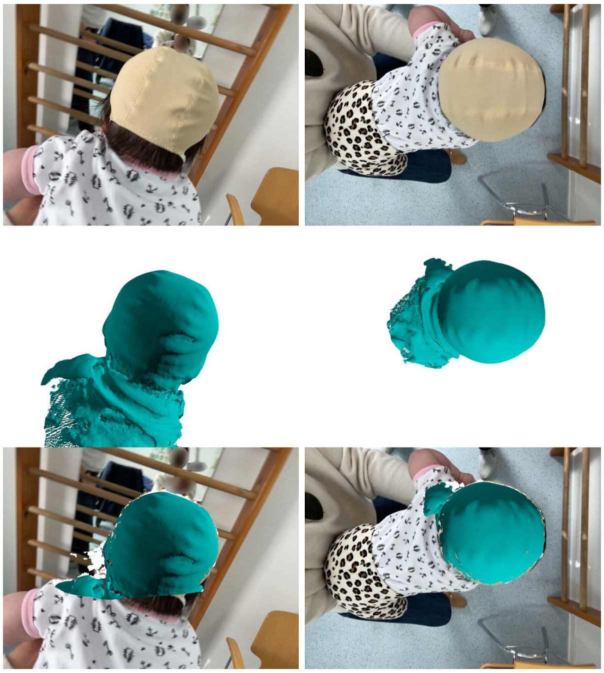
V. 3D RECONSTRUCTION AND MEASUREMENTS

Using the OpenPose keypoints of the ears, the nasion, and the subnasale, the camera trajectory is transformed to conform to



Sample Frame 5 Sample Frame 6

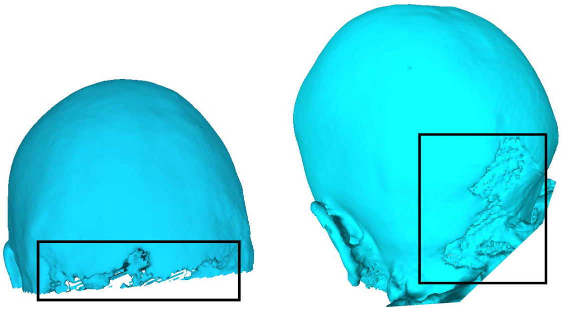
FIGURE 10. **Additional 3D reconstruction results for a second infant. See Fig. 9 for details.**



Sample Frame 7 Sample Frame 8

FIGURE 11. **Additional 3D reconstruction results for a third infant: See Fig. 9 for details.**

the origin and axes described in Section II-A. Missing keypoint detections were annotated manually. Then, all frames that were not discarded during optimization are integrated into a TSDF using the code provided by [40]. The reconstruction volume is discretized with a voxel size of 1 mm. All measured quantities are directly computed using the TSDF. For visualization purposes, the TSDF is converted into a mesh via marching cubes [41].



Reconstruction 4 Reconstruction 5

FIGURE 12. **3D reconstruction artifacts.**

TABLE 1. **Hyperparameter schedule.**

Iteration	λ_{color}	λ_{normal}	λ_{depth}
$1-20$	∞	0.7	$45\,\text{mm}$
$21-40$	∞	0.9	$5\,\text{mm}$
$41-60$	0.1	0.9	$5\,\text{mm}$

The offset of the reference plane is computed with two different methods following [19] (MAX PERIMETER) or [20] (LEVEL3), the corresponding slice is extracted from TSDF volume resulting in a TSDF image. The largest polygon is extracted from the image using marching squares [41] and the perimeter is computed. Then, four line segments LR, AP, DIAG1, and DIAG2 were constructed as depicted in Fig. 1 and Fig. 2.

VI. EVALUATION

A. DESCRIPTIVE STATISTICS

Given a dataset of $n = 8$ samples with caliper measurements $\{y_i\}_{i=1}^n$ and digital measurements $\{f_i\}_{i=1}^n$ of the biparietal diameter (BP), we compute the Pearson correlation coefficient with

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (f_i - \bar{f})^2}}, \tag{14}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \quad \text{and} \quad \bar{f} = \frac{1}{n}\sum_{i=1}^n f_i \tag{15}$$

denote the empirical mean of $\{y_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$ respectively. In the paper, the squared Pearson correlation coefficient r^2 is used.

The measurement deviation is denoted by $c_i = f_i - y_i$ for all $1 \leq i \leq n$. The measurement bias is defined as the empirical mean $\bar{c}$ of $\{c_i\}_{i=1}^n$ with

$$\bar{c} = \frac{1}{n}\sum_{i=1}^n c_i. \tag{16}$$

From the empirical standard deviation σ with

$$\sigma = \frac{1}{n-1}\sum_{i=1}^n (c_i - \bar{c})^2, \tag{17}$$

the estimated standard error of the mean $\hat{\sigma}$ is defined by

$$\hat{\sigma} = \frac{\sigma}{\sqrt{n}}. \tag{18}$$

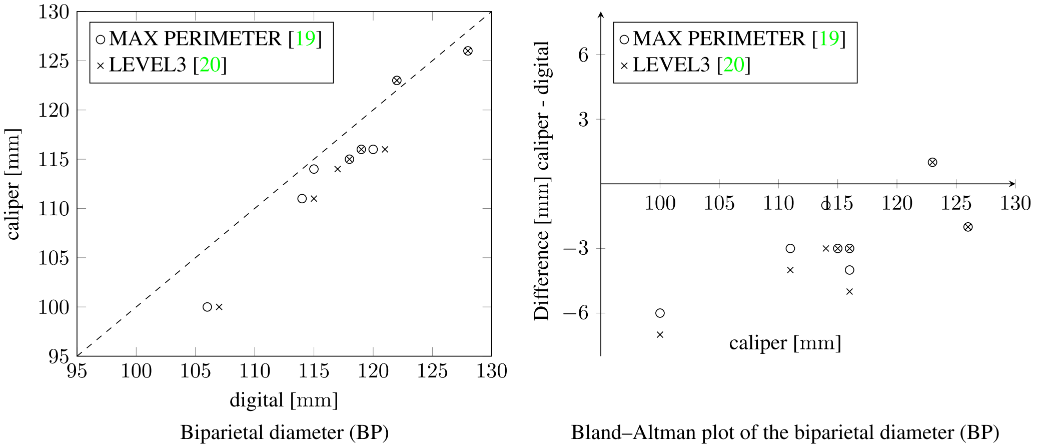


FIGURE 13. Quantitative evaluation on biparietal diameter measured directly with a caliper vs digital measurement on the 3D scans.

We denote the bias in the paper with the pair $(\bar{c}, \hat{\sigma})$ using the notation $\bar{c} \pm \hat{\sigma}$. To evaluate the significance of the estimated digital measurement bias, a t-test is performed with the null hypothesis assuming a bias that is normally distributed with zero mean. The t-value T is then defined by

$$T = \frac{\bar{c}}{\hat{\sigma}} \tag{19}$$

and the respective p-value p following a two-sided test for the null hypothesis is computed via

$$p = 2(1 - F(T, n - 1)), \tag{20}$$

where F denotes the cumulative distribution function of the student's t-distribution evaluated at T with $n - 1$ degrees of freedom.

B. EXPERIMENTAL SETUP

All scans from the created dataset are processed by the proposed pipeline and 3D scans and cranial vault asymmetry measures were extracted. We also experimented with Heges, a commercial RGBD reconstruction app[2] intended to work well on static scenes. Not a single scanning procedure could be completed due to movements of the infant during recording. Commercial solutions requiring remote data processing could not be considered for data privacy concerns.

C. FOREGROUND SEGMENTATION

In Section IV-B and Section IV-E, we propose an application-specific segmentation pipeline. Given the vast literature on trained neural networks for segmentation, it is interesting to evaluate whether such a method could solve the segmentation tasks without additional domain adaptation. To this end, we run the Pyramid Scene Parsing Network [38] on one of our recorded image sequences. The segmentation results are depicted in Fig 7. The predicted pixel classifications indicate that additional domain-specific training data is required to fine-tune these methods to the task at hand and result in usable segmentations. To the best of our knowledge, no infant segmentation dataset is publicly available.

Additionally, we compare the foreground masks estimated by depth thresholding (see Section IV-B) and after refinement (see Section IV-E). Exemplary results are depicted in Fig. 8. We observe an over-segmentation of the head using just the depth thresholding method. Most dynamic pixels are removed from the initial foreground estimation after refinement. Limitations of the proposed method are visible when segmenting the back of the head. The refinement method removes too many pixels. The separation plane estimated in Section IV-E is defined using landmarks in the facial area and near the ears. Additionally, we observe poor detection results of the ears. This error in the plane estimation is more prominent at the back of the head. Natural key landmarks are hard to define and detect at the back of the head.

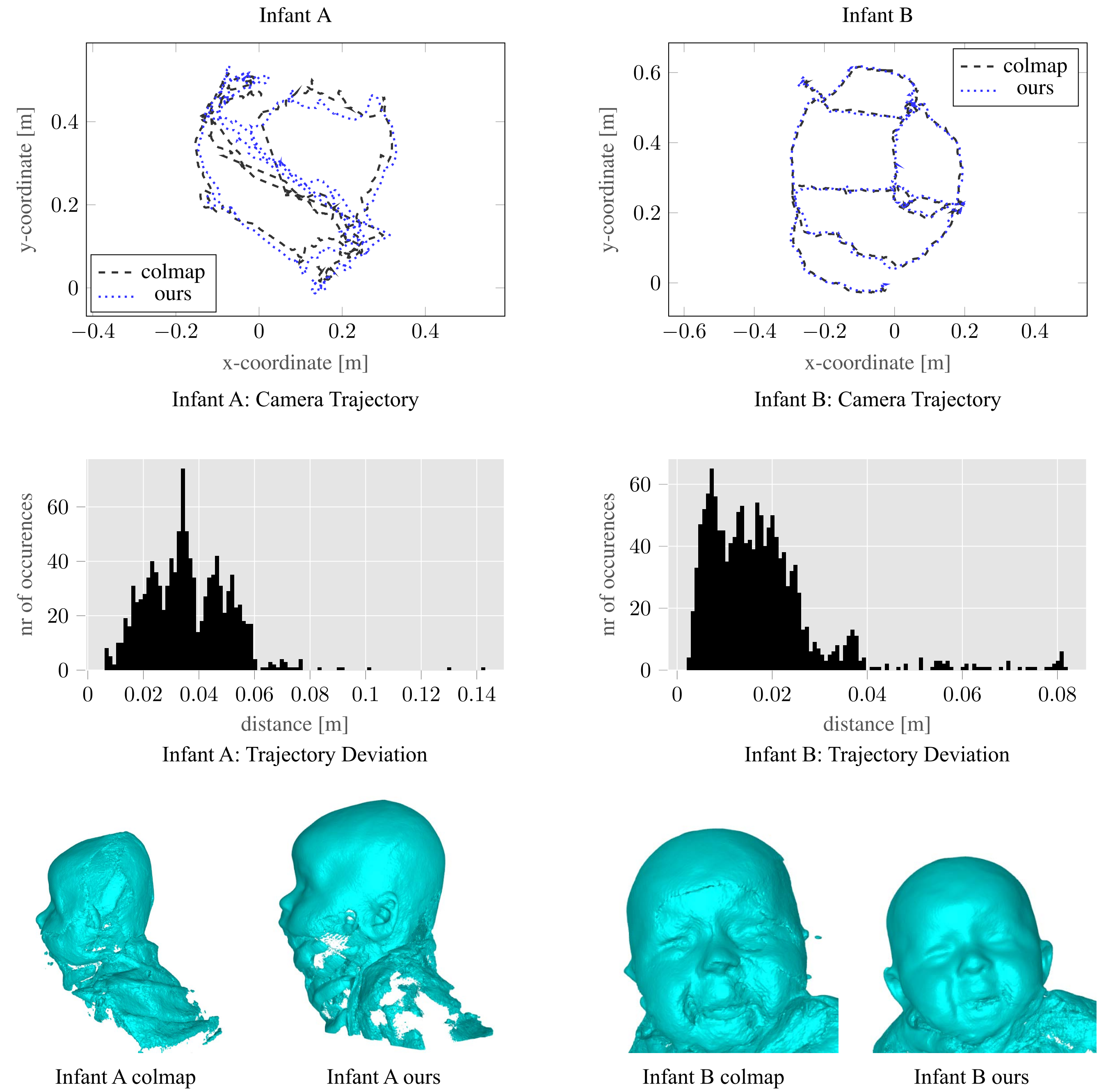
D. QUALITATIVE EVALUATION

We showcase the 3D reconstruction quality and trajectory. The resulting scans and their projection into the source video are depicted in Fig. 9, Fig. 10, and Fig. 11 for a selection of frames showing four different viewing angles each. The complete videos are provided in the supplementary material. Rendering the reconstructed shape over the image sequence gives visually pleasing results. In contrast, 3D reconstruction with Heges was unsuccessful.

The proposed reconstruction pipeline generates 3D meshes even under rapid head movements of the infants. This is possible even if the sensor operator is moving too close to the subject and no depth data is available for multiple frames depicted in the supplementary material.

We also showcase the limitations of our proposed method. In general, visual artifacts are observable in the area at the back of the head, rendering subsequent reliable measurements of the perimeter, AP, DIAG1, and DIAG2 impossible (see Fig. 12). Although the reconstructions are visually pleasing in the face and ear area, small cracks are observed near the back of the head. This is likely due to misaligned camera poses. The foreground mask estimation might be the limiting factor for accurate head reconstruction, consistently oversegmenting or undersegmenting the back of the head

[2] https://apps.apple.com/de/app/heges-3d-scanner/id1382310112



Infant A: Camera Trajectory

Infant B: Camera Trajectory

Infant A: Trajectory Deviation

Infant B: Trajectory Deviation

Infant A colmap

Infant A ours

Infant B colmap

Infant B ours

FIGURE 14. Camera trajectory recovered by colmap and the proposed method. Each row shows the result for a different infant. The first row shows the camera trajectory computed by colmap and the proposed method respectively. Colmap is only using 2D information and a global scale can not be recovered. For this reason, the colmap trajectory is rotated, translated, and scaled to minimize the sum of squared camera distances over all frames. The trajectories are depicted in a top view using orthographic projection. The second row shows two histograms of the distance between the camera positions of colmap and the proposed method over all frames. The last rows shows the reconstructions recovered from our method supplying the camera trajectories computed from colmap and the proposed method respectively. For infant A, the large camera trajectory estimated from colmap results in a failed 3D reconstruction. For infant B, even small deviations in the camera trajectory from colmap lead to clearly visible reconstruction artifacts.

(Fig. 9, Fig. 10, and Fig. 11). OpenPose does not seem to be trained on the viewing angles present in our dataset, leading to infrequent detections of the ears. Training a custom ear detector might alleviate this issue.

E. QUANTITATIVE EVALUATION

We compare our method with colmap [42], [43], an open-source RGB multi-view reconstruction method. Colmap performs 3D reconstruction in two steps: In the first step, the camera poses and a sparse point cloud using SIFT keypoints are recovered [42]. In the second step, a dense point cloud is computed [43]. We run colmap on two sequences with and without incorporation of the initial foreground mask (see Section IV-B). The resulting camera trajectory and 3D reconstructions are depicted in Fig. 14. We proceed by running the multi-view stereo pipeline of colmap, but colmap is unable to construct a dense point cloud from the estimated camera trajectory. Colmap appears to have trouble with the low-textured scene as well as the dynamic scene. These results indicate that an application-specific reconstruction method is required to tackle this problem.

Quantitative evaluation was conducted by comparing the biparietal diameter value measured with a caliper with the measurement extracted from the scans. The measurement plane was extracted from eight scans using the methods MAX PERIMETER and LEVEL3 outlined in Section V resulting in a squared Pearson correlation coefficient of $r^2_{\text{MAX PERIMETER}} = 0.954$ and $r^2_{\text{LEVEL3}} = 0.954$ respectively (see Fig. 13). The bias expressed as the mean deviation and standard error on the mean of both methods are 2.6 ± 0.7 and 3.3 ± 0.8 in millimeters respectively. The p-values for the methods MAX PERIMETER and LEVEL3, 0.009 and 0.005 respectively, suggest a statistically significant observed bias.

Skolnick *et al.* [15] reported an r^2 score of 0.902 and a bias of 4 ± 0.4 on a dataset consisting of 31 scans with extracted measurements. Although our results are competitive, much more samples have to be included in the evaluation and the 3D reconstruction has to be improved, especially in the area at the back of the head.

VII. CONCLUSION

We present a fully automatic 3D infant head reconstruction method using the RGBD stream from a mobile phone. Qualitative and quantitative evaluation suggest accurate 3D reconstruction in the face and ear area bringing us one step closer too an ubiquitous cranial vault asymmetry measuring tool. Future work can address the reconstruction quality in the area near the back of the head, potentially enabling successful extraction of all measurements required for cranial vault asymmetry estimation. Acquiring a bigger dataset would enable an evaluation that highlights the significance of the proposed approach. Additionally, real-time capabilities could enable 3D reconstruction on the mobile device without the need to transfer the data to an external device which could enable real-time feedback to the user improving the camera trajectory.

REFERENCES

[1] AAP Task Force on Infant, "Positioning and SIDS," *Pediatrics*, vol. 89, no. 6, pp. 1120–1126, Jun. 1992. [Online]. Available: https://pediatrics.aappublications.org/content/89/6/1120

[2] R. Y. Moon, R. A. Darnall, L. Feldman-Winter, M. H. Goodstein, and F. R. Hauck, "SIDS and other sleep-related infant deaths: Evidence base for 2016 updated recommendations for a safe infant sleeping environment," *Pediatrics*, vol. 138, no. 5, Nov. 2016. [Online]. Available: https://pediatrics.aappublications.org/content/138/5/e20162940

[3] L. C. Argenta, L. R. David, J. A. Wilson, and W. O. Bell, "An increase in infant cranial deformity with supine sleeping position," *J. Craniofacial Surg.*, vol. 7, no. 1, pp. 5–11, Jan. 1996.

[4] L. A. van Vlimmeren, Y. van der Graaf, M. M. Boere-Boonekamp, M. P. L'Hoir, P. J. M. Helders, and R. H. H. Engelbert, "Risk factors for deformational plagiocephaly at birth and at 7 weeks of age: A prospective cohort study," *Pediatrics*, vol. 119, no. 2, pp. e408–e418, Feb. 2007. [Online]. Available: https://pediatrics.aappublications.org/content/119/2/e408

[5] K. Ditthakasem and J. C. Kolar, "Deformational plagiocephaly: A review," *Pediatric Nursing*, vol. 43, no. 2, p. 59, 2017.

[6] E. Ballardini, M. Sisti, N. Basaglia, M. Benedetto, A. Baldan, C. Borgna-Pignatti, and G. Garani, "Prevalence and characteristics of positional plagiocephaly in healthy full-term infants at 8–12 weeks of life," *Eur. J. Pediatrics*, vol. 177, no. 10, pp. 1547–1554, Oct. 2018.

[7] A. Mawji, A. R. Vollman, J. Hatfield, D. A. McNeil, and R. Sauvé, "The incidence of positional plagiocephaly: A cohort study," *Pediatrics*, vol. 132, no. 2, pp. 298–304, Aug. 2013.

[8] S. Kluba, F. Roßkopf, W. Kraut, J. P. Peters, B. Calgeer, S. Reinert, and M. Krimmel, "Malocclusion in the primary dentition in children with and without deformational plagiocephaly," *Clin. Oral Invest.*, vol. 20, no. 9, pp. 2395–2401, Dec. 2016.

[9] C. Linz, F. Kunz, H. Böhm, and T. Schweitzer, "Positional skull deformities: Etiology, prevention, diagnosis, and treatment," *Deutsches Ärzteblatt Int.*, vol. 114, nos. 31–32, p. 535, 2017.

[10] R. Stücker, "Die mit Plagiozephalus assoziierte Säuglingsasymmetrie," *Zeitschrift Für Orthopädie Unfallchirurgie*, vol. 147, no. 4, pp. 503–512, 2009.

[11] M. Feijen, B. Franssen, N. Vincken, and R. R. W. J. van der Hulst, "Prevalence and consequences of positional plagiocephaly and brachycephaly," *J. Craniofacial Surg.*, vol. 26, no. 8, pp. e770–e773, 2015.

[12] X.-Q. Zhao, L.-Y. Wang, C.-M. Zhao, Q. Men, Z.-F. Wu, and Y.-P. Zhang, "Neurological assessment of Chinese infants with positional plagiocephaly using a Chinese version of the infant neurological international battery (INFANIB)," *Child's Nervous Syst.*, vol. 33, no. 2, pp. 281–288, Feb. 2017.

[13] K. Stoevesandt, H. Ma, U. Beyer, H. Zhang, and G. Jorch, "Positional plagiocephaly in infants," *Monatsschrift Kinderheilkunde*, vol. 166, no. 8, pp. 675–682, 2018.

[14] J. J. Xia, K. A. Kennedy, J. F. Teichgraeber, K. Q. Wu, J. B. Baumgartner, and J. Gateno, "Nonsurgical treatment of deformational plagiocephaly: A systematic review," *Arch. Pediatrics Adolescent Med.*, vol. 162, no. 8, pp. 719–727, 2008.

[15] G. B. Skolnick, S. D. Naidoo, D. C. Nguyen, K. B. Patel, and A. S. Woo, "Comparison of direct and digital measures of cranial vault asymmetry for assessment of plagiocephaly," *J. Craniofacial Surg.*, vol. 26, no. 6, pp. 1900–1903, 2015.

[16] S. Nahles, M. Klein, A. Yacoub, and J. Neyer, "Evaluation of positional plagiocephaly: Conventional anthropometric measurement versus laser scanning method," *J. Cranio-Maxillofacial Surg.*, vol. 46, no. 1, pp. 11–21, Jan. 2018.

[17] A. Breitbarth, T. Schardt, C. Kind, J. Brinkmann, P.-G. Dittrich, and G. Notni, "Measurement accuracy and dependence on external influences of the iphone x truedepth sensor," *Proc. SPIE Photon. Educ. Meas. Sci.*, vol. 11144, Sep. 2019, Art. no. 1114407.

[18] M. Zollhöfer, P. Stotko, A. Görlitz, C. Theobalt, M. Nießner, R. Klein, and A. Kolb, "State of the art on 3D reconstruction with RGB-D cameras," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 625–652, 2018.

[19] P. Meyer-Marcotty, H. Bohm, C. Linz, J. Kochel, A. Stellzig-Eisenhauer, and T. Schweitzer, "Three-dimensional analysis of cranial growth from 6 to 12 months of age," *Eur. J. Orthodontics*, vol. 36, no. 5, pp. 489–496, Oct. 2014.

[20] L. H. Plank, B. Giavedoni, J. R. Lombardo, M. D. Geil, and A. Reisner, "Comparison of infant head shape changes in deformational plagiocephaly following treatment with a cranial remolding orthosis using a noninvasive laser shape digitizer," *J. Craniofacial Surg.*, vol. 17, no. 6, pp. 1084–1091, 2006.

[21] H. F. Jelinek, B. Strachan, B. O'Connor, and A. Khandoker, "A continuous point measure for quantifying skull deformation in medical diagnostics," *Healthcare Technol. Lett.*, vol. 1, no. 2, pp. 56–58, Jun. 2014.

[22] I. Barbero-García, J. L. Lerma, Á. Marqués-Mateu, and P. Miranda, "Low-cost smartphone-based photogrammetry for the analysis of cranial deformation in infants," *World Neurosurg.*, vol. 102, pp. 545–554, Jun. 2017.

[23] M. Tölgyessy, M. Dekan, Ĺ. Chovanec, and P. Hubinský, "Evaluation of the azure Kinect and its comparison to Kinect V1 and Kinect V2," *Sensors*, vol. 21, no. 2, p. 413, Jan. 2021.

[24] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Jun. 2011, pp. 127–136.

[25] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration," *ACM Trans. Graph.*, vol. 36, no. 4, p. 1, Jul. 2017.

[26] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[27] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 343–352.

[28] A. Bozic, M. Zollhofer, C. Theobalt, and M. NieBner, "DeepDeForm: Learning non-rigid RGB-D reconstruction with semi-supervised data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7002–7012.

[29] J. Huang, S.-S. Huang, H. Song, and S.-M. Hu, "DI-fusion: Online implicit 3D reconstruction with deep priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8932–8941.

[30] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Müller-Felber, and A. S. Schroeder, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Int. Conf. Med. Image Comput. Comput. Assist. Intervent (MICCAI)*, Sep. 2018, pp. 792–800.

[31] 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.

[32] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Oct. 2015.

[33] 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.

[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. CVPR*, 2017, pp. 7291–7299.

[35] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. CVPR*, 2017, pp. 1145–1153.

[36] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica, Crystal Phys., Diffraction, Theor. Gen. Crystallogr.*, vol. 34, no. 5, pp. 827–828, 1978.

[37] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. CVPR*, 2017, pp. 2881–2890.

[39] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. (2018). *JAX: Composable Transformations of Python+Numpy Programs*. [Online]. Available: http://github.com/google/jax

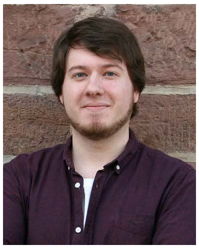
[40] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3DMatch: Learning local geometric descriptors from RGB-D reconstructions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1802–1811.

[41] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, Jul. 1987.

[42] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4104–4113.

[43] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 501–518.

SAMUEL ZEITVOGEL received the B.Sc. and M.Sc. degrees in computer science from the Karlsruhe University of Applied Science, Germany, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer vision with the Chemnitz University of Technology, Germany. Since 2015, he has been with the Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences. His research interests include statistical 3D human shape modeling, computational geometry, large-scale numerical optimization, computer vision, and machine learning.



CHRISTIAN WERNET received the B.Sc. and M.Sc. degrees in computer science from the Karlsruhe University of Applied Sciences, Germany, in 2018 and 2020, respectively. During his bachelor's degree he spent one year as Technical Student at CERN, Geneva, Switzerland, where he helped in the design and implementation of the backend for the new online monitoring system of the CMS experiment. Since 2020, he has been with the Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences. His research interests include computer vision, machine learning, and its applications in industry.

JOHANNES WETZEL received the B.Sc. and M.Sc. degrees in computer science from the Karlsruhe University of Applied Sciences, Germany, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree in computer vision with the Karlsruhe Institute of Technology (KIT), Germany. From 2014 to 2017, he was a Computer Vision Research and Development Engineer with Vitracom AG, Karlsruhe. Since 2017, he has been with the Intelligent Systems Research Group (ISRG), Karlsruhe University of Applied Sciences. His research interests include probabilistic modeling and inference, computer vision, machine learning, and its applications in industry.

ASTRID LAUBENHEIMER received the Diploma degree in mathematics and the Ph.D. degree in 3D model-based computer vision from the University of Karlsruhe, Germany, in 1998 and 2004, respectively. From 2004 to 2006, she was a Postdoctoral Assistant at the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Karlsruhe, Germany. From 2006 to 2009, she was the Head of the Research Group Image-Based Real-Time Systems, Fraunhofer IOSB. Since 2009, she has been a Full Professor in applied computer science at the Karlsruhe University of Applied Sciences, where she is a member and a spokeswoman of the Intelligent Systems Research Group (ISRG). Her research interests include probabilistic modeling and inference, machine learning, and mainly with the applications in computer vision and 3D modeling.

KAI STOEVESANDT received the Diploma degree in physiotherapy from the University of Applied Sciences of Nordhessen, in 2010, and the Diploma degree in osteopathic therapy D.O.T TM and the Diploma degree in osteopathic therapy-pediatrics D.O.T.P TM from the DGOM e.V, in 2015 and 2017, respectively. Since 1996, he has been a Physiotherapist. Since 1997, he has been working at the Ambulanten Zentrum für Rehabilitation und Prävention am Entenfang GmbH (AZR). Since 2007, he has been the Head of the Department of Practice Operations Physiotherapy. Since 2012, he has been actively as a medical device manufacturer of a positioning pillow for infants at VARILAG GmbH & Company KG. Since 2018, he has been the Managing Director of the AZR. His personal focus of work in the therapeutic field is, among other things, the assessment and treatment of infants and children with infantile postural asymmetry and head deformity.