



OPEN Comparison of marker-less 2D image-based methods for infant pose estimation

Lennart Jahn^{1,2}✉, Sarah Flügge¹, Dajie Zhang^{3,4}, Luise Poustka³, Sven Bölte^{5,6,7}, Florentin Wörgötter², Peter B. Marschik^{1,3,4,5} & Tomas Kulvicius^{1,2,3}

In this study we compare the performance of available generic- and specialized infant-pose estimators for a video-based automated general movement assessment (GMA), and the choice of viewing angle for optimal recordings, i.e., conventional diagonal view used in GMA vs. top-down view. We used 4500 annotated video-frames from 75 recordings of infant spontaneous motor functions from 4 to 16 weeks. To determine which pose estimation method and camera angle yield the best pose estimation accuracy on infants in a GMA related setting, the error with respect to human annotations and the percentage of correct key-points (PCK) were computed and compared. The results show that the best performing generic model trained on adults, ViTPose, also performs best on infants. We see no improvement from using specific infant-pose estimators over the generic pose estimators on our infant dataset. However, when retraining a generic model on our data, there is a significant improvement in pose estimation accuracy. This indicates limited generalization capabilities of infant-pose estimators to other infant datasets, meaning that one should be careful when choosing infant pose estimators and using them on infant datasets which they were not trained on. The pose estimation accuracy obtained from the top-down view is significantly better than that obtained from the diagonal view (the standard view for GMA). This suggests that a top-down view should be included in recording setups for automated GMA research.

Keywords Full body pose estimation, Infant motion analysis, Deep neural networks, GMA

Background

When classic biomarker approaches fail to detect developmental conditions early or predict neurodevelopmental outcomes following pre- or perinatal brain lesions or complications during pregnancy, assessments of overt neurofunctions come into play^{1–3}. An early detection of biological markers of adverse outcomes or early diagnosis of developmental conditions facilitates early intervention, when brain plasticity is at its peak, aiming to achieve best possible long-term outcomes⁴. The Prechtl general movement assessment (GMA)⁵ relies on human visual gestalt perception to delineate the spontaneous motor repertoire which is continuously present from fetal life to about 5 months post-term age^{6,7}. The fetus and the newborn move without sensory stimulation in age-specific patterns, so called general movements (GMs). These movements involve the whole body, appear as writhing movements from about 6–9 weeks post-term age and change their character to so-called fidgety movements later on. Neurobiology has coined the term ‘central pattern generator’ (CPG) for the underlying neural circuitry of endogenously generated activity that is not triggered by sensory stimulation but can be modulated by the periphery. General movements that present with a variable sequence of arm, leg, neck, and trunk movements have been studied through human and augmented Gestalt perception approaches^{8–11}. Over the last decades, GMA widened its scope and applicability to serve as a general estimate of the integrity of the developing nervous

¹Child and Adolescent Psychiatry and Psychotherapy, University Medical Center Göttingen; German Center for Child and Adolescent Health (DZKJ), Leibniz Science Campus Göttingen, Von-Siebold-Str. 5, Göttingen, Germany. ²University of Göttingen, III Institute of Physics - Biophysics, Göttingen, Germany. ³Department of Child and Adolescent Psychiatry, Heidelberg University Hospital, Heidelberg University, Heidelberg, Germany. ⁴IDN - Interdisciplinary Developmental Neuroscience, Division of Phoniatrics, Medical University of Graz, Graz, Austria. ⁵Center of Neurodevelopmental Disorders (KIND), Department of Women's and Children's Health, Center for Psychiatry Research, Karolinska Institutet & Region Stockholm, Stockholm, Sweden. ⁶Child and Adolescent Psychiatry, Stockholm Health Care Services, Region Stockholm, Stockholm, Sweden. ⁷Curtin Autism Research Group, Curtin School of Allied Health, Curtin University, Perth, Australia. ✉email: lennart.jahn@phys.uni-goettingen.de

system. It became a tool to look beyond the mere distinction of typical development vs. high risk for cerebral palsy, e.g., to determine the impact of viral infections on postnatal neurodevelopment or describe the early phases of neurodevelopmental conditions such as autism^{12,13}. This broad applicability was the reason we used the multi-camera recording technique for a population based study which included the standard GMA perspective as a use case^{10,13,14}. GMA is a gestalt-based observational tool to classify spontaneous infant motor functions in the first months of life⁵. Infants up to 20 weeks of post-term age are positioned in supine position in a cot and their spontaneous movements, i.e., unstimulated, are video-recorded. These movements can be analyzed and classified into physiological or atypical movement patterns^{2,3,5,6}. Given its high sensitivity and specificity GMA has become one tool of choice for the early detection of cerebral palsy in the early postnatal period³.

State of the art

Recently, efforts have been intensified to automate GMA using computers (for recent reviews see^{8,9,15–18}). In early studies, spontaneous movements in the first months of life were measured with sensors directly attached to the infant^{19,20}. However, the required procedures and the presence of the sensors could potentially change motion⁹, thus, efforts to augment or replace wearable sensors with non-invasive computer vision based methods have become increasingly popular. Early camera based works used optical flow methods^{21–24} for motion detection. With the advent of neural networks for pose estimation^{25–28} current studies have switched to pose estimation and skeleton keypoints for visual-based movement analysis and classification^{29–37}. At first, most automated GMA studies^{29–31,34} used OpenPose^{26,44,45}. More recent approaches changed to newer pose estimators like EfficientPose³², the adaptable pose estimation framework DeepLabCut³³, or a derivative version of OpenPose fine tuned on infant datasets^{36,37}.

Recent findings revealed ViTPose to be the best performing framework for human pose estimation using the COCO dataset²⁸. However, all current pose estimators with large underlying datasets are trained on adults and then used for infant pose estimation without modification. One of the reasons is that there are no large public infant datasets available. The field of infant pose estimation for clinical applications in general faces a complex data sharing issue which presents a true obstacle for all video-based clinical methods³⁴. Infant specific models^{38–42} are consequently mostly trained and tested on silo-datasets (patient data) and results have not yet been compared on a common test-set. There are comparisons between related model architectures when selecting the best model^{40,41}, but they are all trained on the same data. There is also a comparison of different existing models on infant data⁴³, however, there, no infant data was used in training. Thus, it is unknown if the generalization capabilities of specialized infant pose estimators are sufficient for use on datasets the particular models were not trained on.

Additionally, GMA, in its clinical application, is usually done with single 2D-RGB cameras in a diagonal view^{14,52}. This viewing angle is uncommon for pictures of adults on which generic pose estimators are trained, which might also have effects on the pose estimators' performance on infant datasets.

Contribution

In this work, we compare various infant pose estimators (generic and infant specific) to evaluate the performance increase using state of the art methods. We also analyze the generalization capabilities of infant specific pose estimators on datasets different from training, and quantify the effect of different viewing angles on pose estimation accuracy. This is important, as selection of the best pose estimator will directly influence the quality of any task done on the extracted poses.

Specifically, we address the following four main research questions:

- Which generic pose estimator is best suited for infant pose estimation?
To test the generalization capabilities of generic pose estimation models, and to later compare those to the infant specific models, we first analyzed four different pose estimation models, OpenPose^{26,44,45}, MediaPipe^{46,47}, HRNet^{48,49} and ViTPose²⁸. Specific reasons for the inclusion of all models are given in Section "Pose estimation Frameworks". We find that ViTPose performs best.
- Do other infant pose estimators generalize well enough such that they achieve better accuracy on our dataset, or is specific retraining necessary?
We evaluated the infant-specific pose estimators AggPose⁴⁰ and AGMA-HRNet^{48,41} on our dataset. In a second step, we retrained the best performing generic model ViTPose²⁸ on our dataset and then compared the retrained model to the two specific infant pose estimators. As detailed above, a comparison like this has not yet been done for infants using data from a source different from the one used in training. We find that retraining, as predicted, increases accuracy, but also that the infant-specific models do not generalize well and only perform at or below the level of the best generic model.
- Is there a difference in pose estimation accuracy between viewing angles (classic diagonal GMA-perspective vs. top-down)?
We studied the influence of the viewing angle on the pose estimation methods by splitting the analysis by viewing angles. Since the data used in this study comes from a recording setup with multiple angles (see Figure 1a) for later 3D-reconstruction, we can directly compare the different viewing angles that were commonly used in other works, a diagonal view and a top view (see Figure 1b,c), and quantify their influence on pose estimation accuracy. We find better performance for the top view as compared to the diagonal (GMA preferred) view.
- Is having a dedicated estimator per view angle better than one trained on multiple views?
We specifically retrained ViTPose on single view angles, to test if this further increases the accuracy, or at least leads to equal performance on top-down and diagonal views (for single model, the diagonal view was worse in any situation). We find that angle specific ViTPose models lead to no improvement over a unified model.

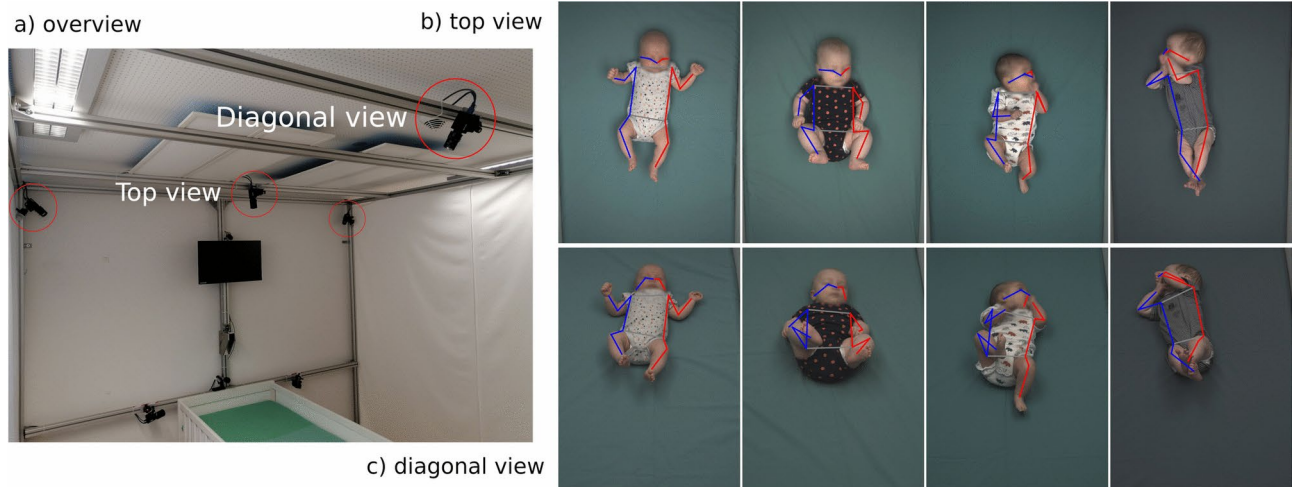


Fig. 1. Overview of the recording setup (a) and its output (b) and (c). The cameras recording the infants are circled in red. For this study, only the two cameras labeled diagonal/top view were used. Panels (b) and (c) show example frames for infants of different age and pose complexity from the two different views. The extracted pose keypoints are displayed as skeletons over the image. Note that neither the human annotators nor any of the pose estimators could reliably determine the position of the fully covered ear in the rightmost example.

Our findings provide useful guidelines for the optimal selection of camera angles and pose estimation models for future projects on automated infant movement classification (such as GMA).

Materials and methods

Dataset

We built a dataset of 4500 frames with COCO⁵³ style labeling (keypoints: nose, eyes, ears, shoulders, elbows, wrists, hips, knees and ankles). The recordings for this analysis were performed utilising a multi-angle marker-less motion tracking setup, specifically designed to record young infants and include 75 recordings of 31 participants from 28 ± 2 days to 112 ± 2 days of gestational age. For this study, we chose the two most common camera angles and relevant view points in the clinical use of GMA: straight down from right above the bed (top-down view) and a view as for a human standing at the foot end of the bed (diagonal view; standard clinical GMA view¹⁴). Figure 1 shows an overview of the camera setup and its output.

Participants

Participants included in this study were prospectively recruited from 2021 to 2023 in Göttingen, Germany, and its close surroundings. The umbrella project aims to investigate cross-domain ontogenetic development in early infancy. To embrace the variability of the targeted dimension, i.e., spontaneous motor functions, 31 participants (17 female) were included. The gestational age at birth of the sample ranged from 34 to 42 weeks. At the time of data analysis, no participant was diagnosed with a neurological or neuromotor deficit, nor any neurodevelopmental impairment. All parents of the participating infants provided written informed consent to study participation and publication of depersonalized data and results. The study was approved by the Institutional Review Board (Ethics Commission) of the University Medical Center Göttingen (19/2019) and performed in accordance with the relevant guidelines and regulations.

Infant movements recordings

Standard laboratory recordings of infant movements at three timepoints (T1 - T3) were included in the current study. These were extracted from data of the umbrella project, assessments, among others, infants' spontaneous movements in a standard laboratory-setting from the 4th to the 18th week of post-term age (PTA; corrected age; from here on, ages refer to the post-term age if not otherwise specified). For the current study, available laboratory recordings ($n = 75$) of the participants for the following timepoints are analyzed: T1: 28 ± 2 days, T2: 84 ± 2 days, and T3: 112 ± 2 days of PTA. As known, infants at 4 weeks (corresponding to T1) present a different spontaneous movement pattern than at 12 and 16 weeks (corresponding to T2 and T3^{7,54}). With standard recordings from T1 to T3, we intended to cover the distinct age-specific spontaneous movement repertoires of the young infants⁶.

Recording Setup

In total, four cameras were mounted to record infants lying in a cot: one from the top down, one diagonally from the foot end and two from the sides. This provides robustness against occlusions and 3D-reconstruction capabilities. The angles were fixed and did not change across the entirety of the dataset. The recordings were

done at 1600×1200 pixels with a frame rate of 60 Hz. All cameras had global shutter and were triggered synchronously for optimal 3D-triangulation capabilities.

For this study, we did not yet do 3D-triangulation, but first determined the optimal pose estimation method on the 2D images alone. The setup also records two side views that are not used in standard GMA settings and were not used for computer vision in this study as well.

Image selection and annotation

The 75 recordings were split into two parts of 50 and 25 which were annotated by one research assistant each, using **JARVIS Annotation Tool 1.2**⁵⁵. Additionally, 10% of the dataset, regardless of the split, were double-annotated for human error and individual joint labeling difficulty estimation.

To select the video frames for the dataset from the available recordings, we first selected time ranges where the infants were in a behavioural state suitable for GMA. From those, we extracted 30 frame-sets each, by *k*-means-clustering (with $k = 30$) on sub-sampled versions of the videos. This yielded good coverage of the different body poses displayed by the infants (see Figure 1 panels b) and c)). In summary, we therefore had 30 frame-sets each obtained from 75 recordings, from which we took two different perspectives. In total, this corresponds to $30 \cdot 75 \cdot 2 = 4500$ frames.

It is not always possible to see all of the keypoints in an image. While there are some instances of eyes, shoulders or ankles being completely occluded, most of the occluded keypoints result from a physiological age-specific ATNR pattern; asymmetric tonic neck reflex. Across the whole dataset, 11.2% of ears could not be seen, because the head was turned on its side. Only counting infants of age smaller than 42 days, the percentage of missing ears was 17.2%, compared to 7.5% for older infants.

Some of the pose estimation frameworks we compared support more or fewer keypoints than COCO, but here we only evaluate the COCO-style keypoints. The annotation software is able to use the 3D calibration information and reproject annotated points within the frame sets. So, although we only labeled the top-down and diagonal views, we supplied the side views to the annotators to deal better with occlusions through re-projection of annotations from the side views (if needed).

Pose estimation frameworks

Generic pose estimation

For comparison, we selected four different generic human pose estimation frameworks.

OpenPose^{26,44,45} This was one of the first pose estimation frameworks used for extraction of movements for analysis and classification of infant motor functions^{29,30,56}. We included this framework as a baseline method to evaluate the performance gain of state-of-the-art models. We used the standard body 25 model available from <https://github.com/CMU-Perceptual-Computing-Lab/openpose>.

MediaPipe pose^{46,47} This framework is not primarily intended for accurate pose estimation, but rather for fast inference on mobile devices. It was interesting for our comparison as Google advertises it as suitable for Yoga and fitness applications and included 25000 images of fitness exercises, so we expected it to deal well with different viewing angles and poses. We used version 0.9.2.1.

HRNet^{48,49} The HR in HRNet stands for high resolution. The particular version used (HRNet-w48 available through MMPose⁵⁷, https://mmpose.readthedocs.io/en/latest/model_zoo/body_2d_keypoint.html#cid-hrnet-on-coco) has an input size of 512×512 pixels. Because our videos are of sufficient resolution and adequate quality with low compression artefacts, this could benefit the pose estimation accuracy.

ViTPose²⁸ This model is the current best performing framework for human pose estimation on the COCO test-dev dataset and uses a Vision Transformer architecture instead of the conventional CNN. We used the version available through mmpose⁵⁷ of the “huge” variety (https://mmpose.readthedocs.io/en/latest/model_zoo/body_2d_keypoint.html#topdown-heatmap-vitpose-on-coco).

Infant pose estimation

We also selected two pose estimation frameworks which were trained on infant images and tested them against generic frameworks. In addition, we also included a ViTPose model which was retrained on our infant dataset (see Section “Retraining”).

AggPose⁴⁰ The AggPose model is based on the Transformer architecture, like ViTPose, and is specifically trained for infant pose detection on a proprietary dataset. In general, the images in the dataset also stem from recordings used for GMA, with infants in supine position, like in our dataset. We used the version published by the authors on github (<https://github.com/PediaMedAI/AggPose>).

AGMA-HRNet48⁴¹ Soualmi et al. also retrained infant pose estimation networks on a proprietary dataset of GMA recordings. Compared to our dataset, the images contain clinical equipment (e.g., tubes, electrodes, cables), leading to more visual clutter in the scenes. The authors retrained multiple networks, of which we chose the HRNet48, because it is the same architecture as the one we used for the comparison of the generic pose estimation frameworks. We used the version published by the authors (<https://drive.google.com/drive/folders/1SEuTqrNdZ6ubRGwMaUazil0BVOqf2cYw?usp=sharing>).

Retraining

We retrained the best performing framework ViTPose (see Section "Generic pose estimators") on our labeled dataset using 5-fold cross-validation. We manually split the dataset such that all images of each infant were contained in one fold only. This ensures the networks do not get better test results by overfitting on specific infants, which are correlated across different recording dates. Each of the five splits therefore contained 900 frames, made up from 450 different situations recorded from two different angles. Because of the different numbers of recordings done for each infant, four splits contained images from six different infants and one split contained images from seven different infants. For training, we randomly split 10% of the training set off as validation set, yielding a final training set size of 3240 frames in each fold. For each fold, we then retrained the model using the default training parameters for ViTPose from the MMPose framework⁵⁷. Our graphics card did not have enough memory to retrain ViTPose-huge, so we retrained the smaller version ViTPose-large instead. We used the pretrained ViTPose model weights for initialization and employed a validation stop on PCK. We then kept the model with the highest PCK on the validation set for evaluation on the test set. The fold with the longest retraining took only 49 epochs until validation stop. Compared to the training of 210 epochs on the substantially larger COCO dataset that went into the base model, this is a relatively inexpensive retraining. All five models are published on Zenodo (<https://doi.org/10.5281/zenodo.14833182>).

The results presented for "Retrained ViTPose" are the combined test set estimations of the five individual models, which in total yield the full dataset.

Quantification metrics

We used two main evaluation metrics to assess the performance of the pose estimation methods: the difference between predicted keypoint position and the human annotation, d_a , in pixels, and the percentage of correct keypoints (PCK).

Our dataset is not biased to any specific side (left or right) of the infants. Moreover, labeling left and right sides of hands, knees, etc., is not inherently different. Therefore, we merged the evaluation results for keypoints of the left and right side variety into one.

Pose estimation error

The difference between model prediction and the human annotation (d_a) is the Euclidean distance between the predicted and human label in pixels:

$$d_a = \sqrt{(p_x - h_x)^2 + (p_y - h_y)^2}, \quad (1)$$

where $p_{x/y}$ are the network predictions and $h_{x/y}$ the human labels for a keypoint.

Percentage of Correct Keypoints (PCK)

For PCK⁵⁸, all keypoint predictions are classified as either correct or incorrect, based on their distance to the label (as given above). The PCK score is then defined as the fraction of correct keypoints out of all keypoints. The distance threshold to be considered correct is defined as a percentage of the apparent torso length of the infant in the frame, measured as the distance between the left shoulder and hip. In this work we considered the PCK at 5%, 7.5% and 10% of the torso size, denoted by PCK@0.05, PCK@0.75 and PCK@0.1, respectively.

Relation to real-world distance

It is not possible to transform any of the quantification measures directly into real-world distances without using 3D triangulation, which was not done in this work, or assuming the points are on the same plane, which is untrue because the infants frequently lift their extremities. For better intuition, we can still provide an upper limit estimation (i.e., the real world distance can not be greater than this) to relate our evaluation metrics to the real-world distance in our recording setup. In our setting, the size of one pixel on the plane of the bed, *always* corresponds to a distance ≤ 0.8 mm (in any camera view). Therefore, 80% of the pixel difference is an upper bound to the real-world distance in mm. Considering the mean torso length of 306 pixels, PCK@0.1 yields the percentage of points that were detected within approximately 2.5 cm of the ground truth.

Although both metrics cannot be directly used for real-world-distances, they are fully appropriate for relative comparisons between the pose estimation frameworks.

Statistical tests

When comparing the results with respect to the difference d_a , we tested the statistical significance of model differences with a paired sample t-test on all keypoints in the dataset (or sub-sets of keypoints if indicated). For PCK, where each point is either correct or incorrect, we used a Pearson's chi-squared test on the frequencies of the outcomes (correct detection or not) to test if the samples for each pose estimation model could come from the same distribution.

Results

Variability in human labeling

Figure 2 shows the mean pixel difference between the two annotators for each keypoint. The results show that the keypoints that were hardest to label are the hips and shoulders.

When comparing the two viewing angles, there is a difference between the diagonal and top-down views. The error for the diagonal view is consistently higher than for the top view. Except for the nose keypoint, the

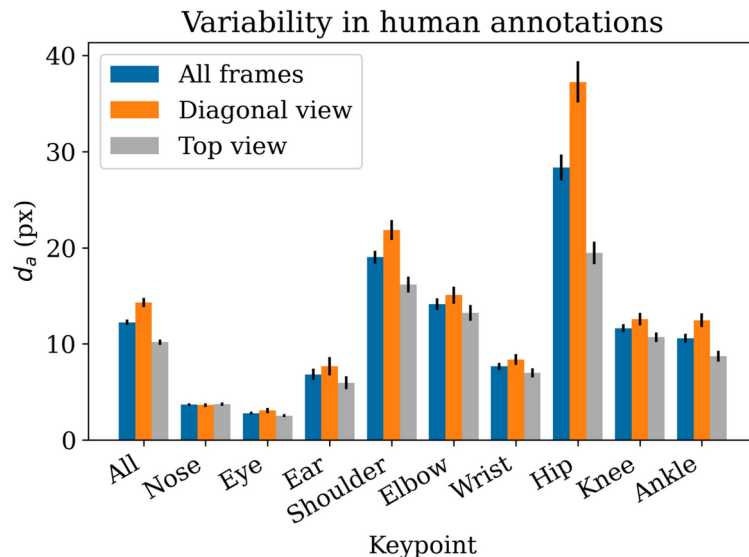


Fig. 2. Difference between two annotators, additionally split by viewing angle. Error bars represent confidence intervals of mean (95%).

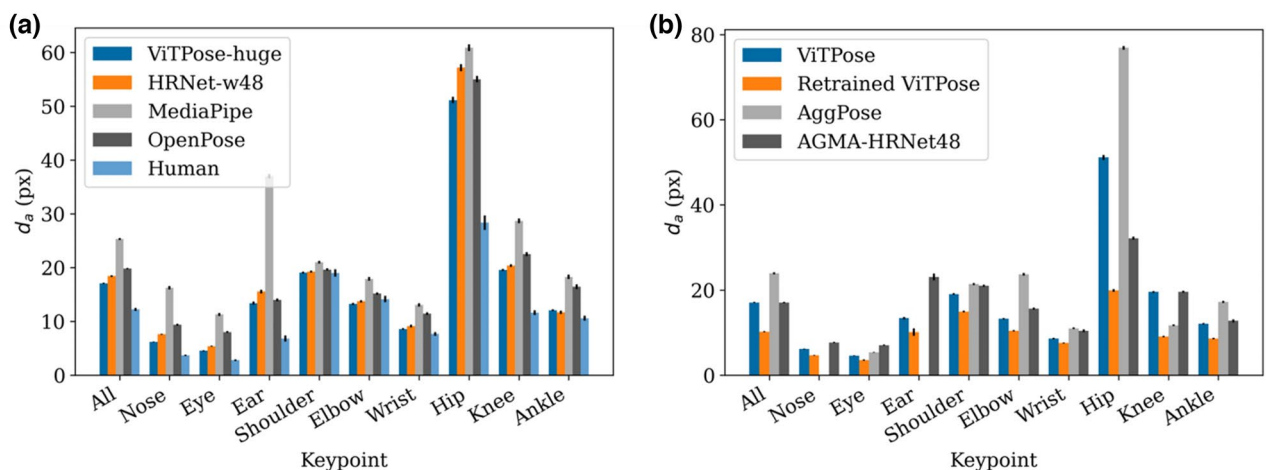


Fig. 3. Difference to annotation d_a in pixels for different subjects, evaluated on our dataset and grouped by key point. Error bars represent confidence intervals of mean (95%). (a) Mean d_a of the different generic frameworks, human for comparison. (b) Mean d_a for the infant pose estimators. Generic ViTPose is also shown for comparison to non-retrained results. Note that AggPose does not estimate the positions of nose and ears.

confidence intervals of the means never overlap. For the hip keypoint, which has the highest error overall, the relative difference between top and diagonal view is also highest.

Generic pose estimators

All four generic pose estimation models were used to estimate keypoints in the whole dataset. The difference between the predictions and the ground truth is shown in Figure 3a. The human labeling difference as presented in Section "Variability in human labeling" is added for comparison. Table 1 shows the corresponding PCK values.

The qualitative result is very close to the difference between human annotators, with the hips having the highest error. But in contrast to the human annotations, the shoulder error is relatively lower, compared to the hip error. Because of the size of the dataset, the differences between the pose estimation models are all statistically significant with $p < 0.001$ (paired sample t-test on all keypoints).

The current state of the art model on the COCO dataset, ViTPose-huge, is also the best performing model on our dataset. However, the relative distance to the second best, HRNet-w48, is only 2.4% for PCK@0.1. For high accuracy requirements, e.g., PCK@0.05, the relative distance is only 0.6%, but the difference is still statistically significant with $p = 0.0312$ (Pearson's chi-squared test). OpenPose, although being an older model,

Model	PCK@0.1	PCK@0.075	PCK@0.05
ViTPose-huge	84.6	75.49	59.50
HRNet-w48	82.6	73.92	59.12
MediaPipe	70.97	58.91	39.99
OpenPose	79.48	70.15	53.74
Retrained ViTPose	93.89	89.82	79.64
AggPose	75.60	67.02	52.43
AGMA-HRNet48	84.32	75.12	59.41

Table 1. PCK values for pose estimation with the different models for the complete dataset. Upper section: generic models, lower section: retrained models. Values of best performing models of their class (generic/specific) are in bold.

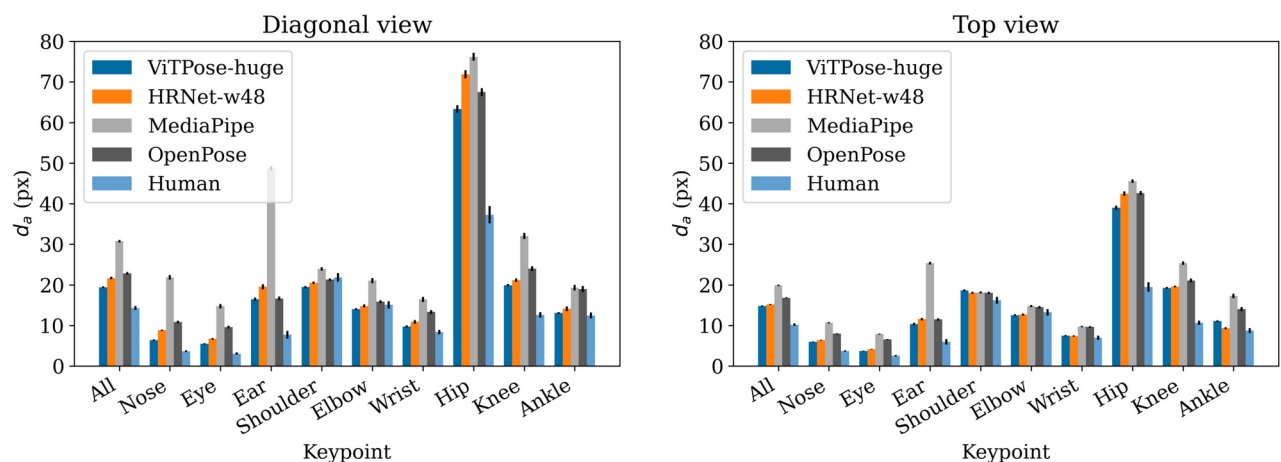


Fig. 4. Pose estimation errors for the generic pose estimation models, split by viewing angle. Error bars represent confidence intervals of mean (95%).

still achieved better results than MediaPipe, which did never achieve a mean error of less than 10 pixels, not even for the clearly defined keypoints like eye and nose.

Infant pose estimators

We evaluated the retrained pose estimation models in the same way we evaluated the generic ones. Figure 3b shows the position estimation errors for the different models with the non-retrained ViTPose network for comparison. The corresponding overall PCK results are displayed in Table 1.

Retraining significantly improved the ViTPose model, with the PCK increasing by 20 percentage points in PCK@0.05 ($p < 0.001$, Pearson's chi-squared test). The mean difference to annotation for the hips is decreased by 61%. All other keypoints also showed significant improvements (all $p < 0.001$, individual paired sample t-tests on keypoints).

The other two infant pose estimators AggPose⁴⁰ and AGMA-HRNet48⁴¹ exhibit different behaviour. Both perform significantly worse than our retrained network ($p < 0.001$, paired sample t-test on all keypoints), with AggPose even performing worse than the non-retrained ViTPose ($p < 0.001$, paired sample t-test on all keypoints). The difference between ViTPose and AGMA-HRNet48 is not statistically significant ($p = 0.66$ for d_a , paired sample t-test on all keypoints, and $p = 0.59$ for PCK@0.05, Pearson's chi-squared test). Regarding the hips, AGMA-HRNet48 improves the accuracy over generic pose estimation, while AggPose is significantly worse (both $p < 0.001$, paired sample t-test on all keypoints).

Influence of the view angle

Figure 4 shows the pose estimation error as Figure 3a, but split into top and diagonal view.

The errors in case of the diagonal view are higher than in the case of the top view. All top-down/diagonal view differences are statistically significant with $p < 0.001$ (paired sample t-test on all keypoints). The most pronounced difference is between the hip position estimation errors. In the top view, the difference is not only lower than in the diagonal view, but also lower as compared to the errors of the other keypoints. This is consistent with the results for the difference between human annotators shown in Figure 2.

MediaPipe does not show any benefit from being designed for fitness applications. In contrary, its performance decreases the most for the diagonal view.

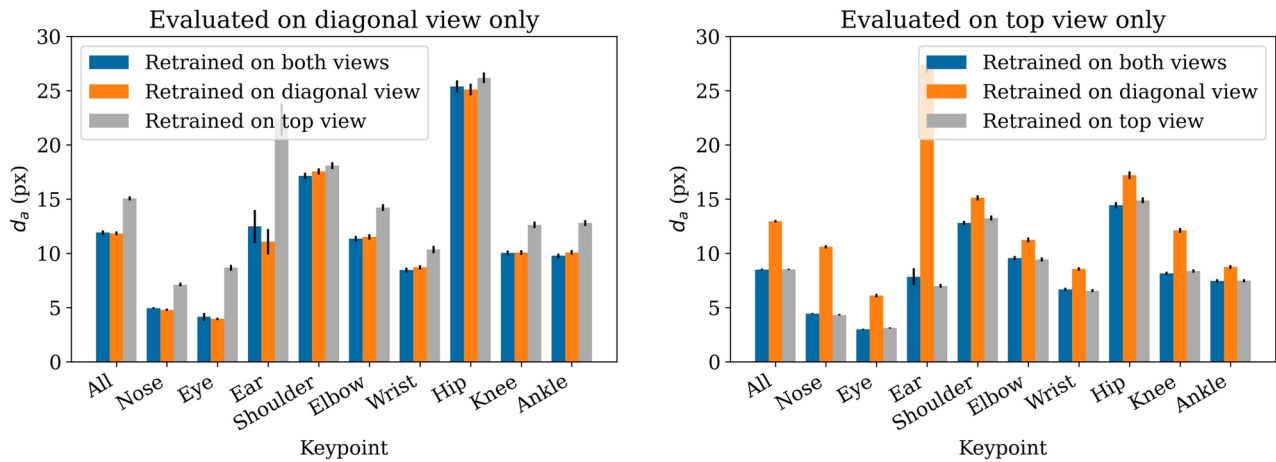


Fig. 5. Mean difference from annotation (d_a) for retrained ViTPose models separately evaluated on diagonal or top view images. Different models have been trained on all, only diagonal view or only top view images, respectively. Error bars represent confidence intervals of mean (95%).

Model	ViTPose-huge	HRNet-w48	MediaPipe	OpenPose	Retrained ViTPose	AggPose	AGMA-HRNet48
Inference speed	1 fps	1.2 fps	3 fps (CPU)	11.5 fps	1.1 fps	2.7 fps	3 fps

Table 2. Inference speed in frames per second (fps), on a NVIDIA GTX 2080 Ti, for each included model. This is the average speed over processing the 4500 frames of the dataset.

Retraining for individual view angles

As was shown in Section "Influence of the view angle", there is a significant difference between the diagonal and top views. We retrained ViTPose models only on the individual viewing angles, to see if specializing on them helps to improve the performance. The results are presented in Figure 5.

There is no significant difference between the model that was trained on both views versus the ones trained on individual views ($p = 0.25$ for diagonal, $p = 0.54$ for top view, both paired sample t-tests on all keypoints), if evaluated against the respective views that were used in training. The results for the respective views that were not used in training, however, are significantly worse ($p < 0.001$, paired sample t-test on all keypoints). This means there is no gain in specializing models to certain viewing angles in our case. However, training on a view that is different from the one used in inference leads to worse performance.

Regarding the general difference in performance between the viewing angles (without specialized models), we again observe worse results in the diagonal view. Even with retraining, it is not possible to achieve the same accuracy of the diagonal view as of the top view.

Inference speed

Since the focus of our study is not on real-time pose estimation for real time applications, we did not optimize the pose estimators (e.g., architectures and/or meta parameters) for inference speed.

All models can be executed on any recent GPU with at least 4GB of memory (for model complexity please see the respective original publications). Table 2 lists the observed inference speed on a NVIDIA GTX 2080 Ti. Those speeds were obtained by using the demonstration scripts of the respective models and adapting them to process our dataset. The results show that all models have acceptable and reasonable inference speed with the ViTPose being the slowest (but the most accurate), and OpenPose the fastest model.

Discussion

There is a wealth of tools for infant movement classification available that are all trained and tested on silo-datasets mostly using one camera view, and thus not directly comparable³⁴. We therefore used a newly designed multi-view dataset that allowed for a direct comparison of the available models.

Generic pose estimators The state of the art models ViTPose and HRNet show clear improvements compared to the older model OpenPose. The gain in PCK@0.05 of 5.76 percentage points is small compared to the gain later achieved by retraining, but still statistically significant ($p < 0.001$, Pearson's chi-squared test).

In addition to the position accuracy, the reliability of keypoint detection also increases with newer models (see Figure S1). A discussion is given in the Appendix.

Infant specific pose estimators The comparison with the other infant pose estimators showed that the generalization capabilities of neural networks are overestimated in this field of study. Our results show that the specialized infant pose estimators do not perform substantially better on our infant dataset than the generic ViTPose model. AGMA-HRNet48 showed no significant overall difference to generic ViTPose, but improved on the hip keypoint. AggPose however, performed worse for all keypoints.

While the AGMA-HRNet48 does improve pose estimation accuracy of the hip, other keypoints get worse, so there is no statistically significant difference between AGMA-HRNet48 and generic ViTPose ($p = 0.66$, paired sample t-test on all keypoints). It has to be noted that the training data of this model and our training data were both derived from a GMA setting. They differ in the amount of visual clutter in the image (compare with Figure 3 in⁴¹), but otherwise closely resemble each other. The infant specific retraining improved the HRNet48 by 0.29 and 1.72 percentage points in PCK@0.05 and PCK@0.1 on our dataset, respectively, putting AGMA-HRNet48 on the same performance level as the generic ViTPose. Note, that the AGMA-HRNet48 model is not the best performing one from its original paper⁴¹. Their DarkPose32 model scored 2.23 and 0.55 percentage points higher than HRNet48 at PCK@0.05 and PCK@0.1, respectively. Since the non-retrained DarkPose32 model performed worse than the non-retrained HRNet48 in⁴¹, we do not expect it to yield a significant improvement over the generic pose estimation on our data either.

AggPose also uses Vision Transformers, like ViTPose, but is trained on a substantially larger dataset than our retrained model (20748 images⁴⁰), still it performs even worse than generic ViTPose, which suggests it is also overfit and cannot sufficiently generalize to our (and potentially to other) infant dataset although the underlying dataset is also from a GMA related setting. AggPose has been compared to other generic pose estimators by Yin et al. (see⁴², Tables 1, 2 and 3), where it also performs worse than the generic pose estimators on different infant pose datasets.

To date, there are no publicly available infant datasets suitable for GMA (mostly due to patient data privacy issues; see³⁴), so we could not evaluate our model on data from a different setup. This is probably further contributing to the general overfitting situation we identified, because there is no training set available that covers multiple GMA related setups. See the Appendix for a comment on a preliminary analysis with the SyRIP dataset³⁹.

Influence of the view angle There is a significant influence of the viewing angle on pose estimation accuracy. Every performance, human, generic and retrained models, is worse on the diagonal view than on the top-down view. We see two main reasons for this.

One reason for the reduced performance is a lack of training data from the corresponding view angle. For everyday situations like those in the COCO dataset, the viewing angle of the diagonal camera corresponds to placing the camera very close to the floor. This is uncommon. We included the MediaPipe framework in our evaluation, because it is marketed towards fitness applications and for the use on mobile devices, a common application in GMA⁵⁹. This suggested to us that it might be trained on images with uncommon view angles (e.g., mobile phone put on the ground to observe the fitness-workout). The training dataset even contained 25000 frames from fitness exercises⁴⁷. However, the model performed worse compared to all others, with no beneficial effect on the diagonal angle. It has to be said though, that MediaPipe is a small model (with respect to the number of parameters) designed for inference on mobile devices and as such not as potent as the other models.

Another reason for the better results of the top view angle is occlusion of features. The anatomy of the baby makes it easy to align the arms and legs with the sight line axis of the camera when lifting them, blocking the view on body keypoints like knee, wrist, shoulder or hip (see the second example from the left in Figure 1b,c). It might be preferable for computer vision to use a diagonal angle from the head end of the bed, instead of the foot end, to make the extremities lift into an orthogonal direction to the camera sight line.

Retraining for individual viewing angles Even when retraining the model to the specific view angle, we could not achieve the same performance on the diagonal view as on the top view (see Figure 5). We suppose this is because the human variation in the training data was still higher in the diagonal view (see Figure 2), because of occlusions. An interesting follow-up would be to only indirectly label both top and diagonal view via reprojection from the side views (which might be hard to do, as they are only the supporting views for occlusion resistance) and see if the significant difference between the view angles persists.

Conclusion

We compared four different generic pose estimation frameworks on our dataset of 4500 frames recorded with a multi-view camera setup designed for automated GMA. The current best performing model on COCO test-dev, ViTPose, also achieved the best results on our infant dataset. Retraining the best performing network on our dataset, expectedly, increased the performance, especially on the keypoint that is most difficult to detect, the hips. In comparison with the other infant pose estimation frameworks, we achieved 20 percentage points better PCK@0.05 than the second best model, AGMA-HRNet48, on our dataset. The other models performed below or on the level of the best generic pose estimator. This suggests that, if possible, one should always retrain on the specific dataset to achieve the best pose estimation accuracy. If retraining is not possible, a state-of-the-art generic pose estimator should be used, unless the dataset is very similar to the training data of a specialized infant pose estimator.

We also compared the pose estimation error between different viewing angles. There is a significant difference in accuracy between the diagonal and top views in both human annotation and neural network pose estimation. Therefore, the results strongly suggest that the top view should be preferred in any setup for *automated* GMA

with only one camera. Since the different viewing angles lead to different accuracy, we tested if specialized pose estimators trained on particular views would improve the accuracy, however, no significant effect was found.

In summary, our study suggests that pose estimation accuracy would benefit from the utility of the top view and the pose estimator being trained on the specific dataset. However, a generic pose estimator should be preferred over a specialized pose estimator trained on a different dataset, if retraining is not feasible. Whether potentially also movement classification accuracy would be improved with more accurate pose estimation remains a question, which needs to be addressed in future work. While the standard GMA method uses a diagonal view for assessment, clinical setups aiming to generate data for AI and computer vision approaches should in future consider top-down viewing angles for recording new data.

Data availability

The authors do not have permission to share the infant video recording data. Derived data, like the annotations and pose estimation data are available from the corresponding author upon reasonable request. The five re-trained ViTPose models are published on Zenodo (<https://doi.org/10.5281/zenodo.14833182>).

Received: 13 December 2024; Accepted: 26 March 2025

Published online: 09 April 2025

References

- Marschik, P. B., Einspieler, C., Sigafoos, J., Enzinger, C. & Bölte, S. The interdisciplinary quest for behavioral biomarkers pinpointing developmental disorders. *Dev Neurorehabil* **19**, 73–74. <https://doi.org/10.3109/17518423.2014.916362> (2016).
- Einspieler, C. et al. The General Movement Optimality Score-Revised (GMOS-R) with Socioeconomically Stratified Percentile Ranks. *J Clin Med* **13**, 2260. <https://doi.org/10.3390/jcm13082260> (2024).
- Novak, I. et al. Early, Accurate Diagnosis and Early Intervention in Cerebral Palsy: Advances in Diagnosis and Treatment. *JAMA Pediatr* **171**, 897. <https://doi.org/10.1001/jamapediatrics.2017.1689> (2017).
- Kirton, A. et al. Perinatal stroke: Mapping and modulating developmental plasticity. *Nat Rev Neurol* **17**, 415–432. <https://doi.org/10.1038/s41582-021-00503-x> (2021).
- Prechtl, H. F. R. et al. An early marker for neurological deficits after perinatal brain lesions. *The Lancet* **349**, 1361–1363 (1997).
- Einspieler, C., Marschik, P. B. & Prechtl, H. F. R. Human motor behavior: Prenatal origin and early postnatal development. *Zeitschrift für Psychologie/Journal of Psychology* **216**, 147–153 (2008).
- Einspieler, C., Prayer, D. & Marschik, P. B. Fetal movements: The origin of human behaviour. *Developmental Medicine & Child Neurology* **63**, 1142–1148. <https://doi.org/10.1111/dmcn.14918> (2021).
- Irshad, M. T., Nisar, M. A., Gouverneur, P., Rapp, M. & Grzegorzec, M. AI Approaches towards Prechtl's Assessment of General Movements: A Systematic Literature Review. *Sensors* **20**, 5321. <https://doi.org/10.3390/s20185321> (2020).
- Silva, N. et al. The future of general movement assessment: The role of computer vision and machine learning - A scoping review. *Research in Developmental Disabilities* **110**, 103854 (2021).
- Marschik-Zhang, D. et al. Building Blocks for Deep Phenotyping in Infancy: A Use Case Comparing Spontaneous Neuromotor Functions in Prader-Willi Syndrome and Cerebral Palsy. *J Clin Med* **12**, 784. <https://doi.org/10.3390/jcm12030784> (2023).
- Kulvicius, T. et al. Deep learning empowered sensor fusion boosts infant movement classification. *Commun Med* **5**, 1–12. <https://doi.org/10.1038/s43856-024-00701-w> (2025).
- Einspieler, C. et al. Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or Rett syndrome. *Research in Autism Spectrum Disorders* **8**, 286–291 (2014).
- Einspieler, C. et al. Association of infants exposed to prenatal Zika virus infection with their clinical, neurologic, and developmental status evaluated via the general movement assessment tool. *JAMA Network Open* **2**, e187235 (2019).
- Einspieler, C., Prechtl, H., Bos, A., Ferrari, F. & Cioni, G. Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants. *Clin. Dev. Med* **167**, 1–91 (2004).
- Raghuram, K. et al. Automated movement recognition to predict motor impairment in high-risk infants: a systematic review of diagnostic test accuracy and meta-analysis. *Developmental Medicine & Child Neurology* **63**, 637–648 (2021).
- Leo, M., Bernava, G. M., Carcagni, P. & Distant, C. Video-based automatic baby motion analysis for early neurological disorder diagnosis: State of the art and future directions. *Sensors* **22**, 866 (2022).
- Peng, Z. et al. Continuous sensing and quantification of body motion in infants: A systematic review. *Heliyon* **9** (2023).
- Reinhart, L. et al. Artificial intelligence in child development monitoring: A systematic review on usage, outcomes and acceptance. *Intelligence-based medicine* 100134 (2024).
- Karch, D. et al. Kinematic assessment of stereotypy in spontaneous movements in infants. *Gait & Posture* **36**, 307–311. <https://doi.org/10.1016/j.gaitpost.2012.03.017> (2012).
- Philippi, H. et al. Computer-based analysis of general movements reveals stereotypes predicting cerebral palsy. *Developmental Medicine & Child Neurology* **56**, 960–967. <https://doi.org/10.1111/dmcn.12477> (2014).
- Adde, L., Helbostad, J. L., Jensenius, A. R., Taraldsen, G. & Støen, R. Using computer-based video analysis in the study of fidgety movements. *Early Human Development* **85**, 541–547 (2009).
- Adde, L. et al. Early prediction of cerebral palsy by computer-based video analysis of general movements: A feasibility study. *Developmental Medicine & Child Neurology* **52**, 773–778 (2010).
- Orlandi, S. et al. Detection of Atypical and Typical Infant Movements using Computer-based Video Analysis. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3598–3601, <https://doi.org/10.1109/EMBC.2018.8513078> (2018).
- Tsuji, T. et al. Markerless Measurement and Evaluation of General Movements in Infants. *Sci Rep* **10**, 1422. <https://doi.org/10.1038/s41598-020-57580-z> (2020).
- Mathis, A. et al. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience* **21**, 1281–1289. <https://doi.org/10.1038/s41593-018-0209-y> (2018).
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 172–186. <https://doi.org/10.1109/TPAMI.2019.2929257> (2021).
- Groos, D., Ramampiaro, H. & Ihlen, E. A. EfficientPose: Scalable single-person pose estimation. *Appl Intell* **51**, 2518–2533. <https://doi.org/10.1007/s10489-020-01918-7> (2021).
- Xu, Y., Zhang, J., Zhang, Q. & Tao, D. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *Advances in Neural Information Processing Systems* **35**, 38571–38584 (2022).
- McCay, K. D. et al. Abnormal Infant Movements Classification With Deep Learning on Pose-Based Features. *IEEE Access* **8**, 51582–51592. <https://doi.org/10.1109/ACCESS.2020.2980269> (2020).
- Reich, S. et al. Novel AI driven approach to classify infant motor functions. *Sci Rep* **11**, 9888. <https://doi.org/10.1038/s41598-021-89347-5> (2021).

31. Nguyen-Thai, B. et al. A spatio-temporal attention-based model for infant movement assessment from videos. *IEEE Journal of Biomedical and Health Informatics* **25**, 3911–3920 (2021).
32. Groos, D. et al. Development and Validation of a Deep Learning Method to Predict Cerebral Palsy From Spontaneous Movements in Infants at High Risk. *JAMA Netw Open* **5**, e2221325. <https://doi.org/10.1001/jamanetworkopen.2022.21325> (2022).
33. Moro, M. et al. A markerless pipeline to analyze spontaneous movements of preterm infants. *Computer Methods and Programs in Biomedicine* **226**, 107119. <https://doi.org/10.1016/j.cmpb.2022.107119> (2022).
34. Marschik, P. B. et al. Open video data sharing in developmental science and clinical practice. *iScience* **26**, 106348. <https://doi.org/10.1016/j.isci.2023.106348> (2023).
35. Gao, Q. et al. Automating general movements assessment with quantitative deep learning to facilitate early screening of cerebral palsy. *Nature Communications* **14**, 8294 (2023).
36. Morais, R. et al. Robust and Interpretable General Movement Assessment Using Fidgety Movement Detection. *IEEE Journal of Biomedical and Health Informatics* **27**, 5042–5053. <https://doi.org/10.1109/JBHI.2023.3299236> (2023).
37. Morais, R. et al. Fine-grained Fidgety Movement Classification using Active Learning. *IEEE Journal of Biomedical and Health Informatics* **1–12**. <https://doi.org/10.1109/JBHI.2024.3473947> (2024).
38. Moccia, S., Migliorelli, L., Carnielli, V. & Frontoni, E. Preterm infants' pose estimation with spatio-temporal features. *IEEE Trans. Biomed. Eng.* **67**, 2370–2380. <https://doi.org/10.1109/TBME.2019.2961448> (2020). [arXiv:2005.08648](https://arxiv.org/abs/2005.08648).
39. Huang, X., Fu, N., Liu, S. & Ostadabbas, S. Invariant Representation Learning for Infant Pose Estimation with Small Data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8. <https://doi.org/10.1109/FG52635.2021.9666956> (IEEE, Jodhpur, India, 2021).
40. Cao, X. et al. AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 5045–5051. <https://doi.org/10.24963/ijcai.2022/700> (2022). [arXiv:2205.05277](https://arxiv.org/abs/2205.05277).
41. Soualmi, A., Ducottet, C., Patural, H., Giraud, A. & Alata, O. A 3D pose estimation framework for preterm infants hospitalized in the Neonatal Unit. *Multimed Tools Appl* <https://doi.org/10.1007/s11042-023-16333-6> (2023).
42. Yin, W. et al. A self-supervised spatio-temporal attention network for video-based 3D infant pose estimation. *Medical Image Analysis* **96**, 103208. <https://doi.org/10.1016/j.media.2024.103208> (2024).
43. Gama, F., Misar, M., Navara, L., Popescu, S. T. & Hoffmann, M. Automatic infant 2D pose estimation from videos: Comparing seven deep neural network methods. <https://doi.org/10.48550/arXiv.2406.17382> (2024). [arXiv:2406.17382](https://arxiv.org/abs/2406.17382).
44. Wei, S.-E., Ramakrishna, V., Kanade, T. & Sheikh, Y. Convolutional pose machines. In *CVPR* (2016).
45. Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR* (2017).
46. Google. MediaPipe - BlazePose. https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker.
47. Bazarevsky, V. et al. BlazePose: On-device Real-time Body Pose tracking. <https://doi.org/10.48550/arXiv.2006.10204> (2020). [arXiv:2006.10204](https://arxiv.org/abs/2006.10204).
48. Wang, J. et al. Deep High-Resolution Representation Learning for Visual Recognition. *TPAMI* (2019).
49. Sun, K., Xiao, B., Liu, D. & Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR* (2019).
50. Blaschek, A. et al. Quantitative Motion Measurements Based on Markerless 3D Full-Body Tracking in Children with SMA Highly Correlate with Standardized Motor Assessments. *J Neuromuscul Dis* **9**, 121–128. <https://doi.org/10.3233/JND-200619> (2022).
51. Zhou, Z. et al. Efficient Domain Adaptation via Generative Prior for 3D Infant Pose Estimation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 51–59. <https://doi.org/10.1109/WACVW60836.2024.00013> (IEEE, Waikoloa, HI, USA, 2024).
52. Marschik, P. B. et al. A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. *Current Neurology and Neuroscience Reports* **17**, 1–15 (2017).
53. Lin, T.-Y. et al. Microsoft COCO: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 (Springer International Publishing, Cham, 2014).
54. Einspieler, C., Peharz, R. & Marschik, P. B. Fidgety movements - tiny in appearance, but huge in impact. *Jornal de Pediatria* **92**, 64–70 (2016).
55. Hüser, T. JARVIS AnnotationTool. <https://github.com/JARVIS-MoCap/JARVIS-AnnotationTool> (2022).
56. Chambers, C. et al. Computer Vision to Automatically Assess Infant Neuromotor Risk. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**, 2431–2442. <https://doi.org/10.1109/TNSRE.2020.3029121> (2020).
57. OpenMMLab. MMPose Pose Estimation Toolbox and Benchmark. <https://github.com/open-mmlab/mmpose> (2020).
58. Yang, Y. & Ramanan, D. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* **35** (2013).
59. Marschik, P. B. et al. Mobile Solutions for Clinical Surveillance and Evaluation in Infancy-General Movement Apps. *JCM* **12**, 3576. <https://doi.org/10.3390/jcm12103576> (2023).

Acknowledgements

We would like to thank the participating families in the SEE-study (Systemic Ethology and Developmental Science) and the additional annotator Felicia Seita; this study was conducted within the SFB 1528, project C03, DFG; clinical aspects are related to VW-IDENTIFIED, BMBF CP-Diadem (13GW0555D) and the DFG P456967546. We were supported by the FWF KLI811. We would like to thank all SEE (Systemic Ethology and Developmental Science, Marschik Lab) members who have contributed to this study.

Author contributions

L.J. P.M. and T.K. developed the idea and research questions. L.J. performed the analysis, on data annotated by S.F., and wrote the main manuscript text. T.K. supervised the work and edited the manuscript from the first draft onward. D.Z., L.P. and P.M. wrote the sections about the clinical aspects and GMA. S.F., D.Z. S.B. F.W. and P.M. edited the manuscript. All authors read and approved the final manuscript version.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests:

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-96206-0>.

Correspondence and requests for materials should be addressed to L.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025