Towards human-level performance on automatic pose estimation of infant spontaneous movements

Daniel Groos [a], Lars Adde [b,d], Ragnhild Støen [b,e], Heri Ramampiaro [c], Espen A.F. Ihlen [a,*]

[a] *Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway*
[b] *Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway*
[c] *Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway*
[d] *Clinic of Clinical Services, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway*
[e] *Department of Neonatology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway*

A B S T R A C T

Assessment of spontaneous movements can predict the long-term developmental disorders in high-risk infants. In order to develop algorithms for automated prediction of later disorders, highly precise localization of segments and joints by infant pose estimation is required. Four types of convolutional neural networks were trained and evaluated on a novel infant pose dataset, covering the large variation in 1424 videos from a clinical international community. The localization performance of the networks was evaluated as the deviation between the estimated keypoint positions and human expert annotations. The computational efficiency was also assessed to determine the feasibility of the neural networks in clinical practice. The best performing neural network had a similar localization error to the inter-rater spread of human expert annotations, while still operating efficiently. Overall, the results of our study show that pose estimation of infant spontaneous movements has a great potential to support research initiatives on early detection of developmental disorders in children with perinatal brain injuries by quantifying infant movements from video recordings with human-level performance.

1. Introduction

During the first months of life, spontaneous infant movements may indicate later developmental disorders, such as cerebral palsy (CP), Rett syndrome, and autism spectrum disorder (Novak et al., 2017; Einspieler et al., 2005, 2014). Early identification of infants at high risk for developmental disorders is essential in order to successfully select appropriate follow-up approaches, and is of greatest importance in research to evaluate early interventions (Støen et al., 2017). The expert-based observation of general movements (GMs) from video recordings, known as the general movement assessment (GMA) (Einspieler et al., 2004), has recently been recommended for clinical use in high-risk infants less than five months of age (Novak et al., 2017). It is especially the fidgety type of GMs, which typically occur between two and five months post-term age, that have shown to predict normal motor development with high accuracy (Einspieler et al., 2016). However, GMA is dependent on individual expert-based training and interpretations, requires time for video observation and analysis, and triggers a high demand for skilled observers if implemented in

large-scale screening (Støen et al., 2017). As an evolving alternative to observational GMA, computer-based methods for objective and consistent risk-assessment are explored (Adde et al., 2010). This supports clinicians in diagnostics, ultimately identifying infants in need for early interventions and focused follow-up care.

Computer-based assessment of infant movements aggregates quantitative movement information from video recordings to yield estimates for the risk of later disorders, like CP (Ihlen et al., 2020). Hence, higher level of correctness in the representation of movement kinematics, such as segment positions and joint angles, facilitates optimal risk analysis. Fidgety movements are small movements of moderate speed and variable acceleration, of neck, trunk, and limbs, in all directions (Einspieler et al., 2004). Automated assessment of such movements requires precise localization of the body parts for proper computer-based risk analysis.

The widespread use of conventional video recordings to capture infant movements has established the need for markerless motion capture, which enables the extraction of movement information in an unobtrusive manner (Rahmati et al., 2015). This provides a low-cost alternative to sensor-based motion capture, which can be performed both at the

---

a) In-Motion Poses   b) Annotated skeleton model

Fig. 1. a) A selection of video frames from In-Motion Poses, originating from standardized and less standardized hospital recordings (top and middle row, respectively), and videos captured from home by parents using the In-Motion smartphone application (Adde et al., 2021) (bottom row). Infant faces are blurred to ensure anonymity. b) The set of 19 body keypoints annotated in the images of In-Motion Poses.

clinic and at home (Adde et al., 2021). Markerless motion capture has the potential to make movement assessments more widely available and promotes worldwide collaboration in analysis of infant movements. Moreover, existing large-scale databases of infant recordings, collected by clinical GMA networks (Støen et al., 2019; Orlandi et al., 2018; Ferrari et al., 2019; Morgan et al., 2019; Kwong et al., 2019; Gima et al., 2019), can be exploited to yield more accurate computer-based methods for risk assessments.

Convolutional neural networks (ConvNets) have improved the techniques for extracting human movement information from conventional 2D videos (Toshev and Szegedy, 2014; Newell et al., 2016; Cao et al., 2019). State-of-the-art markerless motion capture tracks movements automatically through frame-by-frame pose estimation, where the ConvNets predict x and y coordinates of a predefined set of body keypoints, directly from the raw video frames (Andriluka et al., 2014). However, most existing human pose estimation (HPE) methods are targeted towards adults, which compared to infants, differ in anatomical proportions and distribution of body poses (Sciortino et al., 2017). Employed on infant images, the localization performance drops significantly, with 10% of the estimated body keypoint positions placed outside a head length distance from the annotated ground truth positions (i.e., 90% in the PCK_h@1.0 metric described in Section 2.3) (Sciortino et al., 2017). From this, Sciortino et al. (2017) conclude that there is a need to tune HPE ConvNets to the task of infant pose estimation.

Following along these lines, Chambers et al. (2020) retrain the openly available OpenPose network (Cao et al., 2019) by utilizing a dataset of 9039 manually annotated infant images. This improves infant pose estimation, reducing the mean error by 60% (Chambers et al., 2020). Despite this advance, a recent study carried out by our group found that OpenPose lacks the sufficient scaling of network depth, network width, and image resolution for optimal pose estimation (Groos et al., 2020b). Other alternatives to OpenPose, such as DeeperCut (Insafutdinov et al., 2016) used in DeepLabCut (Mathis et al., 2018), posses similar shortcomings as single-scale networks targeted towards multi-person pose estimation. Recent developments in HPE outperform OpenPose and variants by deploying novel multi-scale networks and by maintaining higher spatial resolution (Newell et al., 2016; Sun et al., 2019). OpenPose is also computationally inefficient, which makes it less convenient for real-world applications (Groos et al., 2020b). ConvNet model scaling addresses this challenge by providing trade-offs in localization performance and computational efficiency across various computational budgets (Groos et al., 2020b), better serving single-person applications.

The main objective of the present study is to obtain computationally efficient markerless pose estimation of the spontaneous movements of infants with a localization performance approaching that of human expert annotations. We exploit a large and heterogeneous infant pose dataset covering infant recordings from multiple sites across the world to conduct a comparative analysis of the localization performance and computational efficiency of eight different ConvNet models, including the commonly used OpenPose network. We compare the performance level of the ConvNets with the inter-rater spread of human expert annotations.

2. Materials and methods

In this section, we introduce In-Motion Poses, describe the ConvNet models included in the comparative study, and explain the various performance metrics used to evaluate the ConvNets.

2.1. In-Motion Poses

We developed a dataset comprising infant images with associated human annotations as the ground truth body keypoint positions. We used a large-scale database of 1424 recordings of 9–18 weeks post-term old infants to facilitate pose estimation of the spontaneous movements of infants in supine position across various recording setups. The videos were collected between 2001 and 2018 through different research projects on observational GMA, and all the recordings follow the standards for video-based GMA during the fidgety movement's period (i.e., infants wear a diaper or a onesie, are awake, alert, and content, are not disturbed or using pacifier, and are positioned in the center of a mattress or blanket with the whole body visible) (Einspieler and Prechtl, 2005). The resolution of videos varied from 576×720 to 1080×1920. The study was approved by the regional committee for medical and health research ethics in Norway, under reference numbers 2011/1811 and 2017/913 on 14 January 2019 and 9 October 2019, respectively. Written parental consent was obtained before inclusion.

From these recordings, we proposed a dataset of 20000 video frames. The dataset emphasizes the heterogeneity in spontaneous movements by including videos from 12 different sites from seven countries across the
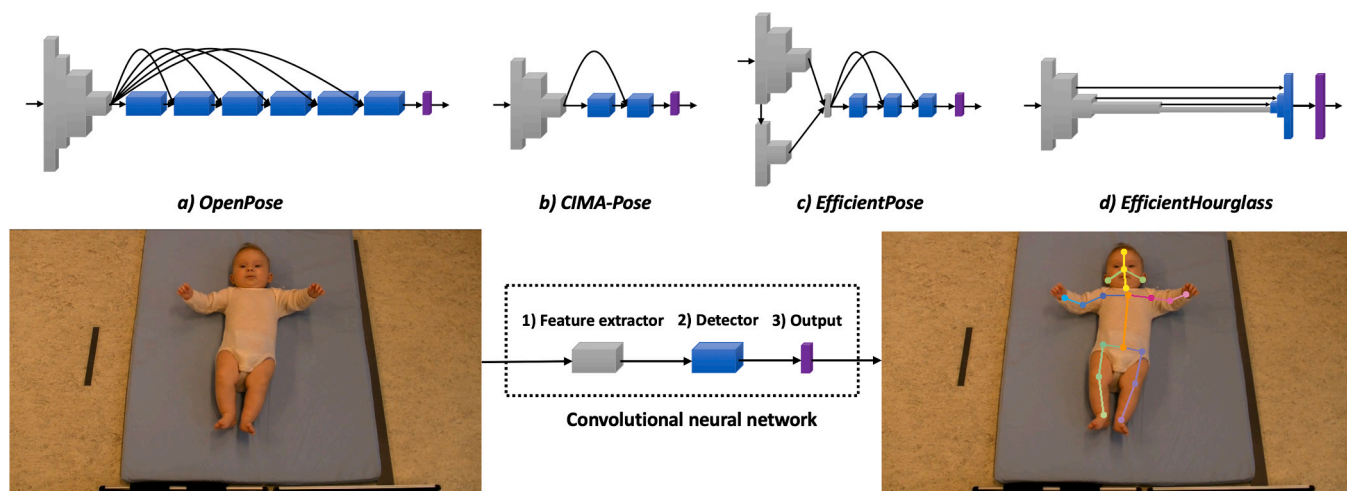
Fig. 2. ConvNets address infant pose estimation from video frames in a frame-by-frame manner by 1) extracting image features, 2) determining features relevant for detection, and 3) estimating infant keypoint positions. The height of the ConvNet blocks (i.e., feature extractor, detector, and output) indicates the block's spatial resolution in relation to the resolution of the input image.

globe (i.e., Norway, India, United States, Turkey, Belgium, Denmark, and Great Britain). The videos cover different groups of infants (e.g., typically developing infants, preterm infants, and other high-risk infants enrolled in hospital-based follow-up programs), and are recorded either by clinicians in a hospital setup or by parents using a smartphone application at home (Adde et al., 2021; Støen et al., 2019) (see Fig. 1a for examples from the dataset). To ensure all video variations were represented, 8000 (40%) frames originated from standardized hospital recordings, 8000 (40%) from home-based smartphone recordings, and the remaining 4000 (20%) from less standardized hospital videos. In each of these three subsets, 80% of the frames were randomly picked with an equal number of frames from each video. Moreover, to achieve proper variation of infant poses, the remaining 20% of frames cover infant poses that occur less frequently, and hence might be particularly challenging for an automatic pose estimator. These frames were manually selected from a random pool of 20000 separate frames (8000, 8000, and 4000 for each subset, respectively), with selection criteria including 1) legs moving towards upper body, 2) overlap of body parts, and 3) crossing of body parts. The resulting total of 20000 frames were split into training (14483 (72%)), validation (1493 (8%)), and test sets (4024 (20%)) in a common machine learning fashion. To mitigate bias and ensure objective evaluation, all frames of a single infant video were placed into one of these three sets.

For the ConvNet models to learn from the data in a supervised fashion, and to be able to validate and test the models, the infant images were annotated to produce the ground truth positions. As depicted by Fig. 1b, 19 distinct body keypoints (i.e., head top, nose, ears, upper neck, shoulders, elbows, wrists, upper chest, right/mid/left pelvis, knees, and ankles) comprised a skeleton model of the infant. The definitions of the body keypoints were agreed upon by a group of human movement scientists and clinical physiotherapists (see Appendix A for a complete overview). Using a separate software tool (Groos and Aurlien, 2018), 10 human expert annotators (two human movement scientists, two physiotherapists, and six engineers) estimated the x and y coordinates of body keypoints, through manual annotation. All body keypoints were annotated in all images regardless of their type of visibility (i.e., visible or occluded). This resulted in a total of 380000 human labels (i.e., 19 annotated keypoint positions for each of the 20000 frames). To measure the consistency between the experts, all annotators estimated the positions of body keypoints in the same sample of 100 randomly selected inter-rater frames. The frames were selected with a similar distribution across recording setups as the full dataset (i.e., 40% standardized, 40% home-based, and 20% less standardized). We computed the inter-rater

annotation disagreement in terms of the mean inter-rater spread H of each body keypoint b. We calculated the mean distance of an annotation $(x_{b,i,j}, y_{b,i,j})$ of an individual expert j of a body keypoint's position in image i, to the average annotation $(\overline{x}_{b,i}, \overline{y}_{b,i})$, across the N (i.e., 10) experts for the S (i.e., 100) frames (see 1). H was normalized according to the head length of the infant in the image, defined as the distance from the top of the head to the upper neck (l_i).

$$H_b = \frac{1}{N \cdot S} \sum_{i=1}^{S} \sum_{j=1}^{N} \frac{\sqrt{(x_{b,i,j} - \overline{x}_{b,i})^2 + (y_{b,i,j} - \overline{y}_{b,i})^2}}{l_i} \qquad (1)$$

2.2. Comparative analysis

By the use of the aforementioned dataset, we trained and evaluated a selection of ConvNet models for the task of infant pose estimation. First, the ConvNet of the state-of-the-art method for infant pose estimation, the OpenPose network (Cao et al., 2019; OpenPose, 2021) (see Fig. 2a for an architectural overview), was trained to yield baseline performance on In-Motion Poses, while also evaluating the official OpenPose library without fine-tuning[1] (OpenPose, 2021). Unless otherwise specified, OpenPose refers to OpenPose ConvNet fine-tuned on In-Motion Poses. Second, we trained a more computationally efficient approach inspired by OpenPose, named CIMA-Pose (see Fig. 2b), which has displayed promising results on infant pose estimation on videos from standardized clinical setups (Groos and Aurlien, 2018). CIMA-Pose comprises a ConvNet with low complexity, reflected by 2.4 million parameters compared to 26 million for OpenPose. OpenPose and CIMA-Pose operate on similar image input resolutions of 368×368 pixels[2]. Third, EfficientPose (Fig. 2c) comprises a family of scalable ConvNets demonstrating 57% improvement in high-precision pose estimation compared to OpenPose, despite significant reduction in computational cost (i.e., FLOPs) and number of parameters (Groos et al., 2020b). EfficientPose yields five model variants, EfficientPose RT and I-IV, obtained by the use of compound model scaling on input resolution, network width, and network depth. The computational requirements of

---

[1] The raw images in In-Motion Poses were downsampled and zero padded to square aspect ratio to achieve the input resolution of the ConvNets.

[2] The latest version of OpenPose (v1.7.0) was used with default settings maintained. Evaluation on In-Motion Poses was performed on the keypoints in the 25-keypoint body model that exist in In-Motion Poses (i.e., all keypoints except head top and upper neck).
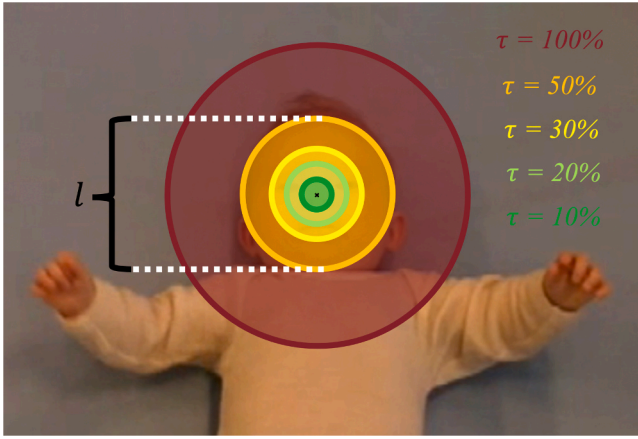
Fig. 3. $PCK_h@\tau$, the percentage of predictions within τl distance from the ground truth location (e.g., nose), is computed across five different thresholds τ (i.e., 100%, 50%, 30%, 20%, and 10%), evaluating the localization performance of a model, from coarse to fine.

EfficientPose span from less than one GFLOP to 74 GFLOPs, which is substantially less than the 161 GFLOPs of OpenPose. Fourth and finally, we optimized an EfficientHourglass model with EfficientNet-B4 backbone (i.e., EfficientHourglass B4) (Groos et al., 2020a), displayed in Fig. 2d. Inspired by the original multi-scale hourglass of Newell et al. (2016), EfficientHourglass performs parallel processing of image features at different scales, while conserving the level of detail (i.e., resolution) inherent in the input image. With an input resolution of 608×608, EfficientHourglass B4 maintains a resolution of at least 152×152 pixels throughout the stages of the network (i.e., feature extractor, detector, and output), compared to the consistent low resolution of 46×46 pixels in the detector and output of the single-scale OpenPose architecture (Cao et al., 2019; Groos et al., 2020a). For further details of the different ConvNets, the reader is referred to their original papers (Cao et al., 2019; Groos et al., 2020a, 2020b; Groos and Aurlien, 2018).

In the experiments, all models (except the underlying model of the official OpenPose library) were trained using a standardized optimization procedure. Pretraining on the general-purpose MPII HPE dataset (Andriluka et al., 2014) was performed, followed by fine-tuning on the training set of In-Motion Poses using the Adam optimizer for 100 epochs with a learning rate of 0.001. We applied data augmentation with random horizontal flipping, scaling (0.75–1.25), and rotation (+/−45 degrees). The optimization procedure was obtained through tuning of models on the validation set of In-Motion Poses.

2.3. Evaluation protocol and performance metrics

To evaluate the localization performance of the models included in the comparative analysis, positions of body keypoints were predicted on the separate test set of In-Motion Poses, comprising 4024 images. The retrained OpenPose, CIMA-Pose, EfficientPose, and EfficientHourglass were evaluated using the model outputs upscaled to input resolution with bilinear interpolation (e.g., three transposed convolutions, each with a stride of 2 and 4×4 kernel, performed $8 \times$ upscaling in Open-Pose, to increase the spatial resolution of outputs from 46×46 to 368×368), omitting the expensive multi-scale testing and flipping procedure commonly used for benchmarking HPE (Tang et al., 2018; Yang et al., 2017), whereas default post-processing was employed with the official version of OpenPose. Model localization performance was determined by comparing the model outputs to human annotations. The performance metrics included percentage of correct keypoints according to head size ($PCK_h@\tau$), normalized mean error (ME), and a proposed metric; percentage of correct keypoints according to human-level

performance ($PCK_h@Human^{0.95}$). $PCK_h@\tau$ computes the fraction of keypoints within τl_i distance from the annotated position, where l_i is the infant head length of image i. To account for both model robustness and performance in high-precision pose estimation, we calculated measures of $PCK_h@\tau$ across various percentages τ of the head size (see Fig. 3). Coarse evaluation was performed with $PCK_h@1.0$, $PCK_h@0.5$, and $PCK_h@0.3$, and fine-grained evaluation by $PCK_h@0.2$ and $PCK_h@0.1$. Moreover, the ME measure reflects the average localization performance of model m on body part b in terms of the mean distance of a model's predictions to the ground truth locations:

$$ME_{m,b} = \frac{1}{S} \sum_{i=1}^{S} d_{m,b,i} \tag{2}$$

where $d_{m,b,i} = \frac{\sqrt{(x_{m,b,i} - \widehat{x}_{b,i})^2 + (y_{m,b,i} - \widehat{y}_{b,i})^2}}{l_i}$ is the Euclidean distance from the estimated keypoint position $(x_{m,b,i}, y_{m,b,i})$ of model m to the human annotation $(\widehat{x}_{b,i}, \widehat{y}_{b,i})$, for keypoint b in image i of the test set. ME was normalized with respect to the head length l_i. To compare model performance against human-level performance, we introduce a metric, called $PCK_h@Human^{0.95}$. $PCK_h@Human^{0.95}$ defines the percentage of model predictions within the 95th percentile of the inter-rater spread of human experts:

$$PCK_h@Human^{0.95}{}_{m,b} = \frac{1}{S} \sum_{i=1}^{S} \delta(d_{m,b,i}) \tag{3}$$

$$\delta(d_{m,b,i}) = \begin{cases} 1, & \text{if } d_{m,b,i} \leq H_b^{0.95} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Here, δ is a binary step function with threshold $H_b^{0.95}$ defining the 95th percentile of the inter-rater spread (where the mean inter-rater spread H_b is specified in Equation 1). In other words, $PCK_h@Human^{0.95}$ is equivalent to $PCK_h@\tau$ when $H_b^{0.95} = \tau$. Thus, $PCK_h@Human^{0.95} = 95\%$ reflects human-level performance. By utilizing the intraclass correlation coefficient (ICC) proposed by Fisher (1992), we also compared consistency (i.e., $ICC(C, 1)$) and agreement (i.e., $ICC(A, 1)$) between model localization error and inter-rater spread across body parts. The ICC values, and associated 95% confidence intervals, between the model ME and the inter-rater spread H of the human experts were calculated using a two-way model. Perfect agreement and consistency with inter-rater spread across body keypoints (i.e., $ICC(A, 1) = ICC(C, 1) = 1$) will suggest that a model displays human-level performance.

In addition to model localization performance, we evaluated the computational efficiency of the ConvNet models. We provide measures for model complexity (number of parameters), computational cost (FLOPs), and inference time (latency). The inference latency per image was estimated from model predictions on an NVIDIA GTX 1080 Ti GPU with TensorFlow 2.5, CUDA 11.0, and CUDNN 8.1. We used a batch size of 128 and computed the median latency in milliseconds over 10 computational runs.

2.4. Sample efficiency

To assess the amount of training data required for ConvNets to converge on the task of infant pose estimation, we carried out experiments with variation in the number of images in the training set, across a range of samples from no fine-tuning[3] to 100 images to the full training set of 14483 infant frames. To evaluate differences in sample efficiency between different ConvNet architectures, experiments were carried out for the most accurate ConvNet in each of the four model families. All experiments were performed over 100 epochs of training, and model

---

[3] When models were evaluated without fine-tuning, predictions were made only on the subset of 16 body keypoints that were available both in the MPII dataset and In-Motion Poses.

Table 1

The performance of the different ConvNets, pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, as well as the official OpenPose library (OpenPose, 2021), in terms of localization performance on the test set of In-Motion Poses, and computational efficiency of the ConvNets from run-time experiments on an NVIDIA GTX 1080 Ti GPU.

Model	Resolution	Localization performance						Computational efficiency		
		@1.0*	@0.5*	@0.3*	@0.2*	@0.1*	ME	Parameters	FLOPs	Latency
OpenPose library	-	96.99%	95.51%	90.90%	81.49%	49.66%	0.1432**	-	-	62.33*** ms
OpenPose	368×368	99.94%	99.61%	97.65%	90.40%	54.89%	0.1087	26,011,743	161,077,013,640	35.21 ms
CIMA-Pose	368×368	99.98%	99.83%	98.74%	93.09%	59.69%	0.0988	2,380,495	15,645,092,494	11.49 ms
EfficientPose RT	224×224	99.96%	99.69%	98.15%	92.15%	58.71%	0.1022	481,336	955,490,248	5.06 ms
EfficientPose I	256×256	99.98%	99.83%	98.81%	93.68%	60.78%	0.0974	743,476	1,785,432,722	7.05 ms
EfficientPose II	368×368	99.97%	99.84%	98.54%	92.41%	62.25%	0.0969	1,759,372	7,944,292,598	19.38 ms
EfficientPose III	480×480	99.99%	99.94%	99.54%	97.57%	78.21%	0.0732	3,258,888	23,777,830,318	41.92 ms
EfficientPose IV	600×600	99.98%	99.93%	99.45%	96.77%	71.10%	0.0834	6,595,430	73,621,311,041	96.48 ms
EfficientHourglass B4	608×608	99.99%	99.95%	99.56%	97.67%	81.11%	0.0681	18,699,936	27,009,544,472	47.01 ms

* PCK$_h$@1.0, PCK$_h$@0.5, PCK$_h$@0.3, PCK$_h$@0.2, and PCK$_h$@0.1 are abbreviated as @1.0, @0.5, @0.3, @0.2, and @0.1, respectively.
** Keypoints in certain images, where the OpenPose library lack predictions due to not being confident, are excluded in computation of *ME*.
*** Latency estimate of the OpenPose library includes time required to pre-process images and perform default post-processing of ConvNet predictions.

Table 2

The localization performance of OpenPose, CIMA-Pose, EfficientPose III, and EfficientHourglass B4, all pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, on the test set of In-Motion Poses, in relation to human-level performance (i.e., inter-rater spread *H*) across body parts *b*, as evaluated by the proposed $PCK_h@Human^{0.95}$ metric.

b	H_b	$H_b^{0.95}$	$PCK_h@Human^{0.95}$			
			OpenPose	CIMA-Pose	EfficientPose III	EfficientHourglass B4
Head top	0.0554	0.1158	60.39%	57.60%	81.59%	89.31%
Nose	0.0301	0.0574	32.03%	42.89%	74.48%	82.41%
Right ear	0.0603	0.1906	88.57%	92.40%	94.41%	92.00%
Left ear	0.0502	0.1364	73.31%	77.49%	88.54%	89.04%
Upper neck	0.0527	0.1212	80.67%	83.23%	88.77%	89.19%
Right shoulder	0.0531	0.1106	62.97%	73.14%	85.71%	86.63%
Right elbow	0.0429	0.0956	52.81%	71.00%	81.71%	86.73%
Right wrist	0.0386	0.0851	45.43%	60.93%	80.14%	82.60%
Upper chest	0.0643	0.1200	69.38%	72.44%	77.31%	79.42%
Left shoulder	0.0576	0.1204	63.25%	60.71%	88.07%	88.74%
Left elbow	0.0418	0.0959	48.19%	46.92%	82.50%	85.69%
Left wrist	0.0388	0.0901	48.83%	52.44%	79.08%	84.74%
Mid pelvis	0.0781	0.1587	82.75%	82.50%	86.43%	90.01%
Right pelvis	0.0812	0.1553	78.31%	80.89%	87.30%	88.72%
Right knee	0.0549	0.1119	66.58%	77.24%	86.63%	89.02%
Right ankle	0.0417	0.0902	51.07%	60.21%	75.47%	80.79%
Left pelvis	0.0828	0.1603	79.25%	77.53%	88.07%	90.31%
Left knee	0.0489	0.1049	49.06%	48.29%	88.22%	89.71%
Left ankle	0.0408	0.0861	45.75%	47.24%	75.70%	82.38%
All body parts	*0.0534*	*0.1161*	*62.03%*	*66.58%*	*81.59%*	*86.71%*

a) OpenPose   *b) CIMA-Pose*   *c) EfficientPose III*   *d) EfficientHourglass B4*   *e) Human annotation*
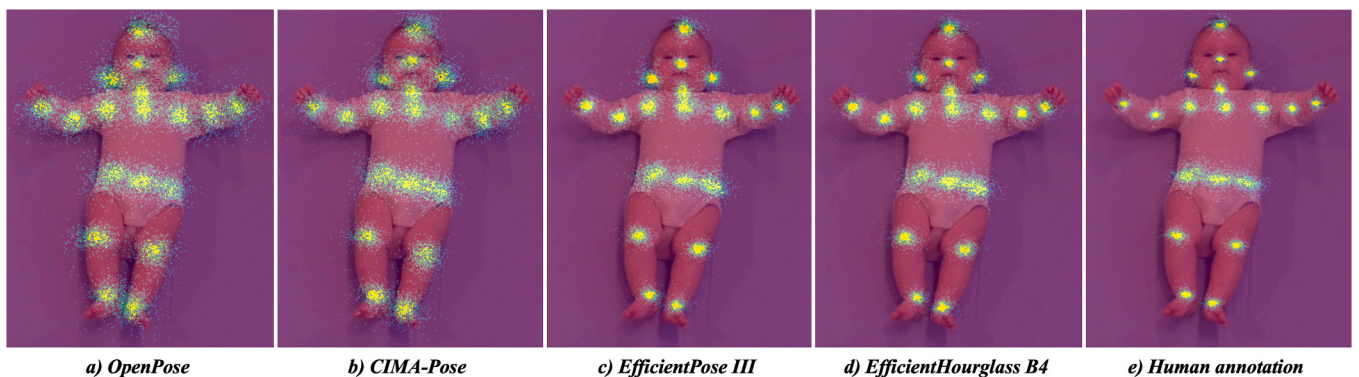
Fig. 4. From left: a-d) The distribution of model prediction errors of the different ConvNets on 1000 randomly sampled frames (according to the distribution of standardized hospital recordings, home-based smartphone recordings, and less standardized hospital recordings) from the test set of In-Motion Poses across body parts, and e) the distribution of the inter-rater spread of the 10 human experts across 100 inter-rater frames (i.e., a total of 1000 annotations). The prediction errors are normalized according to the head size of the infant in the sample image.

Table 3

Absolute agreement and consistency (i.e., $ICC(A, 1)$ and $ICC(C, 1)$) of ConvNets in relation to human expert inter-rater spread across body parts, with 95% confidence intervals in brackets.

	OpenPose	CIMA-Pose	EfficientPose III	EfficientHourglass B4
$ICC(A, 1)$	0.00 [−0.03, 0.07]	0.08 [−0.04, 0.32]	0.47 [−0.03, 0.84]	0.64 [−0.03, 0.91]
$ICC(C, 1)$	0.02 [−0.43, 0.46]	0.45 [0.01, 0.75]	0.94 [0.85, 0.98]	0.96 [0.91, 0.99]

performance in ME, $PCK_h@0.5$, and $PCK_h@0.1$ were calculated on the test set of In-Motion Poses. The smaller training samples were constructed by randomly selecting a subset of frames from the original training set, while maintaining the distribution of videos proposed in Section 2.1. Hence, the smaller samples and the full training set have similar variation in recording setups.

3. Results

Table 1 gives an overview of the performances of the eight different ConvNets, as well as the official version of OpenPose, on In-Motion Poses. In terms of localization performance, a 6–37% decrease in ME compared to the OpenPose baseline is achieved. This is supported by a higher robustness (i.e., gains in $PCK_h@1.0$, $PCK_h@0.5$, and $PCK_h@0.3$). In high-precision pose estimation, $PCK_h@0.1$ from 58.71% to 81.11% can be observed, compared to 54.89% and 49.66% for fine-tuned OpenPose and official OpenPose, respectively. With regards to computational efficiency, all models are smaller, with 1.4–54 times fewer parameters, and require less computation than OpenPose, i.e., 2.2–169 times less FLOPs. Moreover, the most computationally efficient ConvNet, EfficientPose RT, achieved run-time performance of 198 frames per second.

Table 2 displays the localization performance of the top-performing ConvNet of each model family. The most accurate model, EfficientHourglass B4, achieved an ME of 0.0681 compared to the average human inter-rater spread H of 0.0534. This equals an average percentage of human-level performance (i.e., $PCK_h@Human^{0.95}$) of 86.71%, compared to 62.03% for OpenPose. Fig. 4 shows a close resemblance between the spread of the human annotations and the estimates of

EfficientPose III and EfficientHourglass B4 across body keypoints. This resemblance was supported by a significant consistency, $ICC(C, 1)$, and high agreement, $ICC(A, 1)$, between the spread of human expert annotations and the mean error of EfficientPose III and EfficientHourglass B4 (see Table 3). The lower $ICC(A, 1)$ compared to $ICC(C, 1)$ reflects a slightly higher ME for the ConvNet models compared to the inter-rater spread H of the human experts. A similar resemblance with human annotations was not achieved with OpenPose.

Fig. 5 illustrates that fine-tuning significantly improves localization performance of infant pose estimation compared to no fine-tuning (i.e., W/O). Moreover, all ConvNets benefit from increased training set size, especially in terms of the $PCK_h@0.1$ measure (Fig. 5c). However, whereas localization performance of OpenPose and CIMA-Pose saturates at sample sizes beyond 5000 images, EfficientPose III and EfficientHourglass B4 benefit from larger training sets. There is also a tendency that EfficientPose III and EfficientHourglass are more stable across dataset sizes, with a smaller difference in localization performance from 100 to 14483 images, compared to OpenPose and CIMA-Pose.

In Fig. 6, the localization performance of EfficientHourglass B4 is assessed qualitatively by providing model predictions on a selection of challenging images (i.e., less frequently occurring infant poses as described in Section 2.1) in the test set of In-Motion Poses.

4. Discussion

The main objective of the study was to obtain computationally efficient markerless infant pose estimation with a level of localization performance approaching that of human expert annotations. A comparative analysis has showed that performance levels comparable to
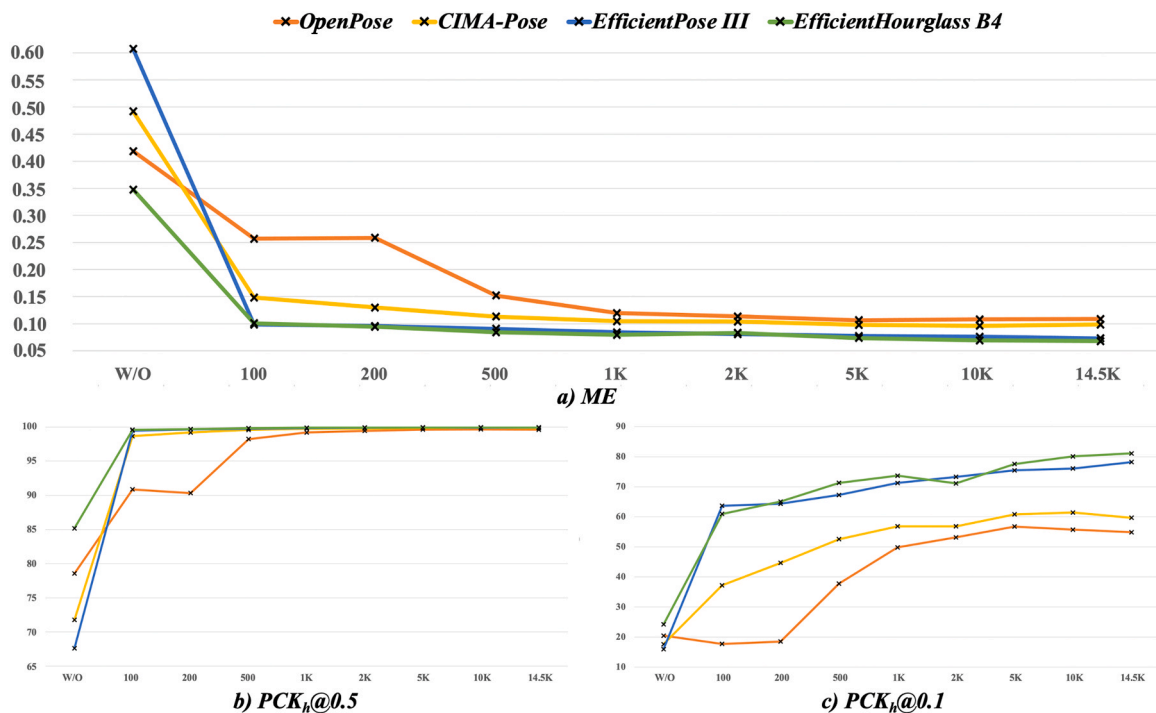

Fig. 5. Localization performance of OpenPose, CIMA-Pose, EfficientPose III, and EfficientHourglass B4, all pretrained on MPII (Andriluka et al., 2014), without fine-tuning (i.e., W/O) and with increasing amounts of data (from 100 to 14483 images) for fine-tuning on In-Motion Poses.

Fig. 6. Predictions of EfficientHourglass B4 on rare but normal infant poses in the test set of In-Motion Poses. The first and second row contain images where the model correctly predicted the position of body keypoints. The third row indicates cases where the model missed certain body keypoints (images from left to right: 1) right ankle, 2) head top and nose, 3) right elbow and right wrist, and 4) right wrist and left wrist). Infant faces are blurred to ensure anonymity.

human expert performance can be achieved, by utilizing contemporary ConvNets for HPE together with an extensive infant video database. This is reflected by $PCK_h@Human^{0.95}$ of the top-performing ConvNets approaching human-level performance, whereas the commonly applied OpenPose network does not reach similar level of localization performance.

4.1. Improving localization performance

The large improvement in localization performance compared to the state-of-the-art method OpenPose (Cao et al., 2019) is due to two main reasons. First, the hypothesis of Sciortino et al. (2017), that HPE ConvNets require fine-tuning on a selection of infant images to perform well on pose estimation of infants, is confirmed. The introduction of a large-scale infant pose dataset, In-Motion Poses, has improved the localization performance of OpenPose from 78.56% to 99.61% on $PCK_h@0.5$, as illustrated by Fig. 5b. Taking into account the error taxonomy of Ruggero Ronchi and Perona (2017), this indicates that the coarse localization errors, like the frequency of inversions (i.e., the predictions that appear at an incorrect body keypoint, such as misinterpretation of the left and right wrist) and misses (i.e., the erroneous localizations that are made without interfering with other keypoints), have been reduced. Despite the increased robustness with regards to coarse prediction errors, the optimal level of localization performance has not been reached. Further improvement of the ConvNets may be achieved by more systematically studying the cases where the models fall short, for example with substantial occlusion of body parts or specific body postures. Fig. 6 indicates that such scenarios exist. Accordingly, we could extend the existing dataset with images that target these situations to further improve model robustness through retraining. In a future perspective, it would also be valuable to assess if we could take into account the temporal information of a video to reduce prediction errors due to occlusion or rare body postures. Pose tracking that extends beyond frame-by-frame pose estimation may achieve this, but current progress in the field is restricted to processing a single pair of video frames with limited gap in time (Bertasius et al., 2019), which may not address cases of prolonged occlusion.

Second, the large improvement in $PCK_h@0.2$, $PCK_h@0.1$, and $PCK_h@Human^{0.95}$ of CIMA-Pose, EfficientPose, and EfficientHourglass

B4, compared to OpenPose, is due to a reduction in fine prediction errors. EfficientPose III, EfficientPose IV,[4] and EfficientHourglass B4 reduce fine prediction errors better than OpenPose by operating on increased input and output resolutions. The consistent high resolution of EfficientHourglass B4 seems to maximize this benefit by displaying the highest values of $PCK_h@0.1$ and $PCK_h@Human^{0.95}$. However, the increase of resolution comes at the cost of reduced computational efficiency, in terms of increased number of FLOPs and decreased latency (see Table 1). Thus, alternative methods for post-processing of ConvNet predictions (e.g., soft-argmax (Levine et al., 2016)), or post-processing of the frame-by-frame position estimates over consecutive frames by low-pass filters, such as median filtering (Tukey, 1977), might reduce fine prediction errors more effectively. However, this demands that the video has a sufficient sample rate (e.g., 60 fps). Furthermore, fine prediction errors may also be minimized by decreasing the spread in annotated keypoint positions. As illustrated in Fig. 4, the distributions of prediction errors of EfficientPose III and EfficientHourglass B4 across body parts resemble the inter-rater spread of the human experts (e.g., higher variation in the placement of the keypoints of the pelvis, compared to the nose keypoint). This indicates that contemporary ConvNets for HPE, when supplied with sufficient amounts of training data (see Fig. 5 for the effect of sample size), are able to maximize the benefit of human annotations. Hence, a hypothesis for further studies is that more precisely annotated keypoints will further eliminate fine prediction errors, by model error being highly correlated with the inter-rater spread of human experts (see Table 3). Consequently, lower variation in the annotation of the keypoints of the pelvis may improve the ability of the ConvNets to localize these keypoints with high localization performance. More consistent annotations between human experts, reflected by lower inter-rater spread, may be obtained by proposing more precise definitions of the keypoint positions, than those in Appendix A. This could be particularly valuable for body keypoints

---

[4] EfficientPose IV displayed lower localization performance than EfficientPose III on In-Motion Poses, due to small batch size during training, which was necessary for the model to fit into GPU memory. As demonstrated by Tables 1 and 5, EfficientPose IV performed better than EfficientPose III in case of similar batch sizes.

that currently have higher inter-rater spread (e.g., for the keypoint of the upper chest). Human expert annotations may also be supplemented or replaced by other methods, such as marker-based solutions and 3D motion capture systems. These approaches may also yield performance improvements beyond fine prediction errors, by providing more precise annotations of occluded keypoints than can be achieved with 2D videos. We suggest that studies on infant pose estimation, and HPE in general (e. g., on challenges such as MPII (Andriluka et al., 2014)), judge localization performance against metrics related to human-level performance, such as $PCK_h@Human^{0.95}$, to evaluate the progress on these tasks in relation to human-level performance.

4.2. Improving computational efficiency

Our comparative analysis has shown that a large model size (i.e., number of parameters) is not necessary for high-precision infant pose estimation. On similar input resolution, both OpenPose and CIMA-Pose were outperformed by the more computationally efficient low-complexity EfficientPose II model on $PCK_h@$ 0.1 (see Table 1). Instead, it appears that high-precision infant pose estimation can be obtained with a relatively small number of parameters. This is demonstrated by EfficientPose III displaying only 5.12% decrease in $PCK_h@Human^{0.95}$, compared to EfficientHourglass B4, despite having 5.7 times fewer parameters. Combining this observation with the influence of high input and output resolution on localization performance, we would suggest further studies to investigate the effect of high resolution with low-complexity ConvNets. This could potentially narrow the current gap in localization performance between computationally efficient ConvNets, such as EfficientPose RT, and high-precision counterparts that are less computationally efficient, like EfficientPose III and EfficientHourglass B4. It would also be of particular interest to systematically study the optimal trade-off between localization performance and computational efficiency, by carefully assessing the localization performance of ConvNets of various complexities across different image resolutions. Our study suggests that ConvNets developed for HPE can be simplified when transferred to the infant pose estimation domain. HPE targets more complex circumstances and environments (e.g., images of multiple persons, a wide range of different activities, individuals of varying age, and substantial occlusion), whereas infant pose estimation is concerned with a single, clearly visible infant in supine position according to the guidelines of GMA (Einspieler et al., 2004; Andriluka et al., 2014). Potential paths for reducing network complexity could be 1) a decrease in network width (i.e., number of feature maps), and 2) less extensive use of multi-scale ConvNet architectures. The former may more appropriately address the little diversity in infant videos compared to the far-reaching HPE task, whereas the latter takes into account the small variation in an infant's distance to the camera and anatomical proportions. Nevertheless, from studying the inference latency of the ConvNets, we observed processing speeds from 10 to 198 fps (Table 1) on an NVIDIA GTX 1080 Ti consumer GPU. Further speedups of the pool of models studied in this paper may be obtained by implementing the ConvNets in low-level code like C++ or CUDA. Thus, a three-minute video of infant spontaneous movements could potentially be processed by a high-precision pose estimator in less than three minutes, which is feasible for clinical use. Moreover, the efficiency of the ConvNets can be further enhanced by utilizing techniques for compressing models with minimal loss of localization performance. Quantization-aware training, knowledge distillation, model pruning, and sparse kernels are paths that are worth to investigate (TensorFlow, 2020; Bucilua et al., 2006; Tung and Mori, 2018; Elsen et al., 2020). By obtaining accelerated and compressed ConvNets, the automatic pose estimation have the potential to be deployed locally at smartphones in the clinic and at home. Thus, infant pose estimation will be more easily applicable, while preserving patient privacy through decentralized processing of infant recordings on local devices.

4.3. External validity

In previous studies on ConvNet-based markerless infant pose estimation from 2D videos, investigations have been restricted to small or synthetic samples of infant videos (Hesse et al., 2018; Chambers et al., 2020). Hence, the external validity of such approaches is debatable, since ConvNets require large amounts of realistic images across various settings related to the task at hand to perform well on pose estimation. In this study, we have utilized a large-scale international database of GMA certified video recordings to train the ConvNets. Subsequently, we have validated the models on a separate set of 284 infant videos from a diverse range of hospital and home-based setups (see Fig. 1a). The high resistance to coarse prediction errors of the evaluated ConvNets suggests that infant pose estimation promotes flexibility in application in real-world scenarios. This encompasses various settings (e.g., clinic, research center, and home), across different countries, and without depending on specific camera equipment. When assessing the transfer validity of the ConvNets fine-tuned on In-Motion Poses on the synthetic dataset proposed by Hesse et al. (2018), only the best performing ConvNet on In-Motion Poses, EfficientHourglass B4, outperformed the official version of the state-of-the-art method OpenPose and displayed an acceptable transfer by maintaining a high level of localization performance (Table 6 and Fig. 8). This could suggest that the high-capacity multi-scale feature extractor of EfficientHourglass B4, through pre-training on MPII (Andriluka et al., 2014) and fine-tuning on In-Motion Poses, has learnt features that generalize beyond the natural infant images of In-Motion Poses. On the contrary, the feature extractors of OpenPose, CIMA-Pose, and EfficientPose are of lower relative capacity and contain fewer abstraction levels (i.e., scales) compared to EfficientHourglass B4 (Fig. 2). Hence, these fine-tuned ConvNets might lack the ability for appropriate transfer beyond recording setups of In-Motion Poses (e.g., plain backgrounds, and natural lighting and shading). However, the consistent localization performance of the official OpenPose library (OpenPose, 2021) (Tables 1 and 6) suggests that training on a sufficiently heterogeneous and large-scale human pose dataset, such as COCO (Lin et al., 2014) of 250000 human poses from various contexts, may combat the lack of high-capacity and multi-scale feature extraction to yield better generalizability. Similar effects could be achieved by combining In-Motion Poses with synthetic or natural infant pose datasets covering the variation in recording setups we want ConvNets to be tuned towards. Nevertheless, we should take into consideration the overall model capacity (i.e., number of parameters), which for CIMA-Pose and EfficientPose might not be sufficient to achieve appropriate transfer from In-Motion Poses to synthetic infants. We could therefore investigate ConvNet compound scaling on infant pose estimation, to determine the appropriate scaling factors of input resolution, network width, and network depth. Further studies should also more thoroughly assess the external validity of the trained ConvNets on real-life infant recordings, to verify that the high level of localization performance demonstrated by the present study indeed can be reproduced. This involves assessing the robustness in operating on video recordings from different recording setups with large variations in aspects, such as video quality, background environment, camera angle, and lighting conditions. The infant pose estimators could also be validated across groups of infants with different age, size, skin color, clothing, and postural variability within datasets like In-Motion Poses. Moreover, the degree of localization performance of the ConvNets in relation to state-of-the-art marker-based motion capture systems could also be assessed (Vicon, 2020; Qualisys, 2020). It is worth stressing that it is unrealistic to expect flawless pose estimation in recording situations highly dissimilar to the settings the models have been trained and evaluated in. However, the models can be retrained on other video databases when keypoint annotations are available. It is also worth investigating if the predefined set of body keypoints is sufficient for performing relevant assessments of characteristics of infant spontaneous movements identified in clinical GMA. However, for applications

emphasizing movement kinematics of other body keypoints (e.g., rotation of hands and feet, and relative movements of fingers or toes), the proposed infant pose estimation can be extended through retraining of ConvNets on different annotated sets of keypoints.

In summary, with improved ConvNet architectures and an extensive database of infant video recordings, body keypoint positions can be estimated with human-level performance. This will enable capturing more subtle infant movements and postures, and, consequently, improve early detection of risk-related infant movement kinematics (Ihlen et al., 2020; Einspieler et al., 2019). These improved ConvNets will also facilitate the assessments of infant movement kinematics which require a high level of detail, like fidgety movements or postural patterns in specific parts of the body, such as side-to-side head movements and atypical head centering (Einspieler et al., 2019).

5. Conclusions

The present study represents a significant progress towards clinically feasible markerless pose estimation of infant movements between two to five months of post-term age. This has been achieved by combining state-of-the-art ConvNets for human pose estimation with a novel heterogenous infant dataset. Highly precise detection of body keypoints enables accurate localization of segments and joints, which may facilitate computer-based assessment of characteristics of infant spontaneous movements related to risk of developmental disorders. With no dependency to body-worn markers, sensors or other expensive laboratory equipment, the automatic infant pose estimation can handle videos both captured by parents at home and by physicians at a hospital clinic. In conclusion, this technology has the potential to facilitate further research initiatives on infant movement analysis and motivate national and worldwide collaborations.

CRediT authorship contribution statement

Daniel Groos: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Lars Adde:** Conceptualization, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Ragnhild Støen:** Conceptualization, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Heri Ramampiaro:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration. **Espen A.F. Ihlen:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Keypoint definitions

The set of 19 body keypoints along with their definitions (see Fig. 7 and Table 4) were agreed upon by an expert group of human movement scientists and infant physiotherapists. The body keypoints were selected to cover most effectively the many degrees of freedom in the infant movements, while at the same time being properly defined to facilitate consistent annotation across humans.
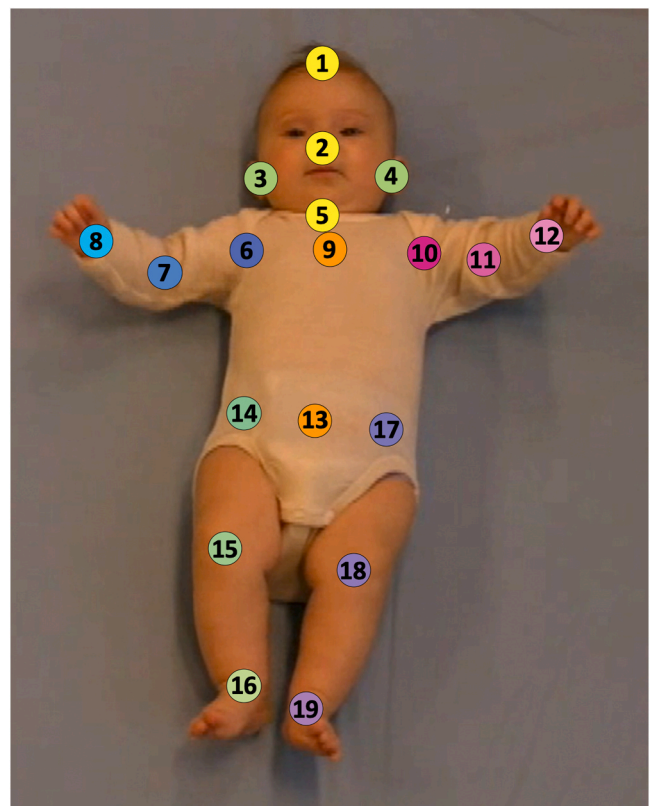

Fig. 7. The placements of the 19 different body keypoints on an infant.

Table 4
Definitions of body keypoints.

#	Body keypoint	Definition
1	Head top	Top of the forehead
2	Nose	Tip of the nose
3	Right ear	Center of the right ear
4	Left ear	Center of the left ear
5	Upper neck	Center of the larynx
6	Right shoulder	Center of the right shoulder joint
7	Right elbow	Center of the right elbow joint
8	Right wrist	Center of the right wrist joint
9	Upper chest	Midway between 6 and 10
10	Left shoulder	Center of the left shoulder joint
11	Left elbow	Center of the left elbow joint
12	Left wrist	Center of the left wrist joint
13	Mid pelvis	Midway between 14 and 17
14	Right pelvis	Right spina iliaca anterior superior
15	Right knee	Center of the right knee joint
16	Right ankle	Center of the right ankle joint
17	Left pelvis	Left spina iliaca anterior superior
18	Left knee	Center of the left knee joint
19	Left ankle	Center of the left ankle joint

Appendix B. Batch size inspection

We assessed the effect of fine-tuning the EfficientPose models on a reduced batch size of four images (i.e., the batch size of EfficientPose IV) to investigate possible performance degrade with EfficientPose IV due to inappropriate batch size. In comparison to Table 1, Table 5 displays performance degrade from training with reduced batch size, most evident in terms of high-precision localization, with 11.20–30.83% reduction in PCK_h@0.1.

Appendix C. Transfer validity

To investigate the transfer validity of the methods in our comparative analysis, we evaluated the localization performance of the models fine-tuned on In-Motion Poses, as well as the official OpenPose library, on the openly available MINI-RGBD dataset proposed by Hesse et al. (2018) (Table 6). The MINI-RGBD dataset comprises 12 synthetic infant video recordings of quite different nature than the recordings in In-Motion Poses. Localization performance, in terms of PCK_h@1.0, PCK_h@0.5, PCK_h@0.3, PCK_h@0.2, PCK_h@0.1, and ME, was measured on the subset of 12 body keypoints that are similar for MINI-RGBD and In-Motion Poses (i.e., nose, upper neck, shoulders, elbows, wrists, knees, and ankles). Since MINI-RGBD does not contain a keypoint for the top of the forehead, the head length of an infant was estimated as two times[5] the distance between the annotated keypoints of the nose and upper neck. This ensures that the evaluation metrics reflect a similar level of correctness as the metrics used with the evaluation on In-Motion Poses in Table 1.

Furthermore, for the most accurate ConvNet, namely EfficientHourglass B4, we conducted a qualitative experiment by estimating the locations of the 19 body keypoints in In-Motion Poses on a randomly selected frame in each of the 12 infant videos in the MINI-RGBD dataset (Fig. 8).

We also supply as Supplementary material frame-by-frame predictions of keypoint locations in a real, external infant recording for the best performing ConvNet in each model family, as well as by the use of the official version of OpenPose. The recording follows the standards for GMA (Einspieler and Prechtl, 2005), and has been recorded using the setup of the In-Motion App (Adde et al., 2021), which is similar to the home-based smartphone recordings in In-Motion Poses.

Table 5
The localization performance of EfficientPose RT and I-III on the test set of In-Motion Poses, when trained with the batch size of EfficientPose IV, followed by the performance difference in relation to the experiments in Table 1.

Model	PCK$_h$@ 1.0	PCK$_h$@ 0.5	PCK$_h$@ 0.3	PCK$_h$@ 0.2	PCK$_h$@ 0.1	ME
EfficientPose RT	99.80% (−0.16%)	99.32% (−0.37%)	92.93% (−5.22%)	72.50% (−19.65%)	27.88% (−30.83%)	0.1717 (0.0695)
EfficientPose I	99.94% (−0.04%)	99.66% (−0.17%)	97.22% (−1.59%)	85.42% (−8.26%)	38.38% (−22.40%)	0.1311 (0.0336)
EfficientPose II	99.98% (0.01%)	99.78% (−0.06%)	98.01% (−0.53%)	89.85% (−2.56%)	49.73% (−12.52%)	0.1137 (0.0168)
EfficientPose III	99.99% (0.00%)	99.94% (0.00%)	99.47% (−0.07%)	96.48% (−1.09%)	67.01% (−11.20%)	0.0884 (0.0152)

Table 6
The transfer validity of the different ConvNets, pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, and the official OpenPose library (OpenPose, 2021), in terms of localization performance on the MINI-RGBD dataset (Hesse et al., 2018).

Model	PCK$_h$@1.0	PCK$_h$@0.5	PCK$_h$@0.3	PCK$_h$@0.2	PCK$_h$@0.1	ME
OpenPose library	98.35%	97.02%	94.47%	90.75%	73.80%	0.1030
OpenPose	88.59%	79.59%	71.77%	62.27%	38.41%	0.3926
CIMA-Pose	95.72%	88.99%	81.27%	71.83%	46.68%	0.2415
EfficientPose RT	94.98%	91.28%	86.91%	79.98%	53.83%	0.2135
EfficientPose I	93.13%	91.09%	88.16%	81.98%	56.19%	0.2772
EfficientPose II	92.49%	90.41%	87.41%	80.57%	54.60%	0.3263
EfficientPose III	83.79%	81.45%	79.60%	76.06%	58.56%	0.8559
EfficientPose IV	93.02%	91.15%	89.05%	86.14%	71.35%	0.2565
EfficientHourglass B4	99.81%	99.17%	97.52%	94.13%	75.86%	0.0845

---

[5] The head length of an infant (i.e., the distance from head top to upper neck) in In-Motion Poses was in average 1.98 times the distance from nose to upper neck.

Fig. 8. Predictions of keypoint locations of EfficientHourglass B4 in randomly selected frames from videos in the MINI-RGBD dataset (Hesse et al., 2018).

Appendix D. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compmedimag.2021.102012.

References

Adde, L., Brown, A., Van den Broeck, C., De Coen, K., Horsberg Eriksen, B., Fjørtoft, T., Groos, D., Ihlen, E.A., Osland, S., Pascal, A., et al., 2021. The In-Motion-App for remote general movement assessment: a multi-site observational study. BMJ Open 11. https://doi.org/10.1136/bmjopen-2020-042147.

Adde, L., Helbostad, J.L., Jensenius, A.R., Taraldsen, G., Grunewaldt, K.H., Støen, R., 2010. Early prediction of cerebral palsy by computer-based video analysis of general movements: a feasibility study. Dev. Med. Child Neurol. 52, 773–778.

Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B., 2014. 2D human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686–3693.

Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J., Torresani, L., 2019. Learning temporal pose estimation from sparsely-labeled videos. Adv. Neural Inf. Process. Syst. 3027–3038.

Bucilua, C., Caruana, R., Niculescu-Mizil, A., 2006. Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535–541.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A., 2019. OpenPose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell.

Chambers, C., Seethapathi, N., Saluja, R., Loeb, H., Pierce, S.R., Bogen, D.K., Prosser, L., Johnson, M.J., Kording, K.P., 2020. Computer vision to automatically assess infant neuromotor risk. IEEE Trans. Neural Syst. Rehabil. Eng. 28, 2431–2442.

Einspieler, C., Bos, A.F., Krieber-Tomantschger, M., Alvarado, E., Barbosa, V.M., Bertoncelli, N., Burger, M., Chorna, O., DelSecco, S., DeRegnier, R.A., et al., 2019. Cerebral palsy: early markers of clinical phenotype and functional outcome. J. Clin. Med. 8, 1616.

Einspieler, C., Kerr, A.M., Prechtl, H.F., 2005. Is the early development of girls with rett disorder really normal? Pediatr. Res. 57, 696–700.

Einspieler, C., Peharz, R., Marschik, P.B., 2016. Fidgety movements-tiny in appearance, but huge in impact. J. De. Pediatr. 92, S64–S70.

Einspieler, C., Prechtl, H.F., 2005. Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. Ment. Retard. Dev. Disabil. Res. Rev. 11, 61–67.

Einspieler, C., Prechtl, H.R., Bos, A., Ferrari, F., Cioni, G., 2004. Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. Mac Keith Press.

Einspieler, C., Sigafoos, J., Bartl-Pokorny, K.D., Landa, R., Marschik, P.B., Bölte, S., 2014. Highlighting the first 5 months of life: General movements in infants later diagnosed with autism spectrum disorder or rett syndrome. Res. Autism Spectr. Disord. 8, 286–291.

Elsen, E., Dukhan, M., Gale, T., Simonyan, K., 2020. Fast sparse ConvNets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14629–14638.

Ferrari, F., Plessi, C., Lucaccioni, L., Bertoncelli, N., Bedetti, L., Ori, L., Berardi, A., DellaCasa, E., Iughetti, L., D'Amico, R., 2019. Motor and postural patterns concomitant with general movements are associated with cerebral palsy at term and fidgety age in preterm infants. J. Clin. Med. 8, 1189.

Fisher, R.A., 1992. Statistical methods for research workers. In: Breakthroughs in Statistics. Springer, pp. 66–70.

Gima, H., Shimatani, K., Nakano, H., Watanabe, H., Taga, G., 2019. Evaluation of fidgety movements of infants based on gestalt perception reflects differences in limb movement trajectory curvature. Phys. Ther. 99, 701–710.

Groos, D., Adde, L., Ihlen, E.A., 2020a. Approaching human precision on automatic markerless tracking of human movements. Gait Posture 81, 117–118.

Groos, D., Aurlien, K., 2018. Infant body part tracking in videos using deep learning – facilitating early detection of cerebral palsy (Master's thesis), NTNU.

Groos, D., Ramampiaro, H., Ihlen, E.A., 2020b. Efficientpose: scalable single-person pose estimation. Appl. Intell. 51 (4), 2518–2533. https://doi.org/10.1007/s10489-020-01918-7.

Hesse, N., Bodensteiner, C., Arens, M., Hofmann, U.G., Weinberger, R., Sebastian Schroeder, A., 2018. Computer vision for medical infant motion analysis: state of the art and RGB-D data set. In: Proceedings of the European Conference on Computer Vision (ECCV).

Ihlen, E.A., Støen, R., Boswell, L., de Regnier, R.A., Fjørtoft, T., Gaebler-Spira, D., Labori, C., Loennecken, M.C., Msall, M.E., Möinichen, U.I., et al., 2020. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: a multi-site cohort study. J. Clin. Med. 9, 5.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B., 2016. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. Springer, pp. 34–50.

Kwong, A.K., Eeles, A.L., Olsen, J.E., Cheong, J.L., Doyle, L.W., Spittle, A.J., 2019. The baby moves smartphone app for general movements assessment: engagement amongst extremely preterm and term-born infants in a state-wide geographical study. J. Paediatr. Child Health 55, 548–554.

Levine, S., Finn, C., Darrell, T., Abbeel, P., 2016. End-to-end training of deep visuomotor policies. J. Mach. Learn. Res. 17, 1334–1373.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. 21, 1281–1289.

Morgan, C., Romeo, D.M., Chorna, O., Novak, I., Galea, C., DelSecco, S., Guzzetta, A., 2019. The pooled diagnostic accuracy of neuroimaging, general movements, and neurological examination for diagnosing cerebral palsy early in high-risk infants: a case control study. J. Clin. Med. 8, 1879.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer, pp. 483–499.

Novak, I., Morgan, C., Adde, L., Blackman, J., Boyd, R.N., Brunstrom-Hernandez, J., Cioni, G., Damiano, D., Darrah, J., Eliasson, A.C., et al., 2017. Early, accurate diagnosis and early intervention in cerebral palsy: advances in diagnosis and treatment. JAMA Pediatr. 171, 897–907.

OpenPose, 2021. Real-time multi-person keypoint detection library for body, face, hands, and foot estimation. ⟨https://github.com/CMU-Perceptual-Computing-Lab/openpose⟩ (Accessed 30 May 2021).

Orlandi, S., Raghuram, K., Smith, C.R., Mansueto, D., Church, P., Shah, V., Luther, M., Chau, T., 2018. Detection of atypical and typical infant movements using computer-based video analysis. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 3598–3601.

Qualisys, 2020. Human biomechanics.⟨https://www.qualisys.com/applications/human-biomechanics/⟩ (Accessed 17 July 2020).

Rahmati, H., Dragon, R., Aamo, O.M., Adde, L., Stavdahl, Ø., Van Gool, L., 2015. Weakly supervised motion segmentation with particle matching. Comput. Vis. Image Underst. 140, 30–42.

Ruggero Ronchi, M., Perona, P., 2017. Benchmarking and error diagnosis in multi-instance pose estimation. Proc. IEEE Int. Conf. Comput. Vis. 369–378.

Sciortino, G., Farinella, G.M., Battiato, S., Leo, M., Distante, C., 2017. On the estimation of children's poses. In: International Conference on Image Analysis and Processing. Springer, pp. 410–421.

Støen, R., Boswell, L., De Regnier, R.A., Fjørtoft, T., Gaebler-Spira, D., Ihlen, E.A., Labori, C., Loennecken, M., Msall, M., Möinichen, U.I., et al., 2019. The predictive accuracy of the general movement assessment for cerebral palsy: a prospective, observational study of high-risk infants in a clinical follow-up setting. J. Clin. Med. 8, 1790.

Støen, R., Songstad, N.T., Silberg, I.E., Fjørtoft, T., Jensenius, A.R., Adde, L., 2017. Computer-based video analysis identifies infants with absence of fidgety movements. Pediatr. Res. 82, 665–670.

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5693–5703.

Tang, W., Yu, P., Wu, Y., 2018. Deeply learned compositional models for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 190–206.

TensorFlow, 2020. Quantization aware training. https://www.tensorflow.org/model_optimization/guide/quantization/training (Accessed 17 April 2020).

Toshev, A., Szegedy, C., 2014. DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1653–1660.

Tukey, J.W., 1977. Exploratory data analysis, 2, Reading, MA.

Tung, F., Mori, G., 2018. CLIP-Q: deep network compression learning by in-parallel pruning-quantization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7873–7882.

Vicon, 2020. The most trusted mocap ecosystem. https://www.vicon.com/applications/life-sciences/ (Accessed 17 July 2020).

Yang, W., Li, S., Ouyang, W., Li, H., Wang, X., 2017. Learning feature pyramids for human pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp. 1281–1290.

Daniel Groos: Daniel Groos received the M.Sc. degree in Computer Science from the Norwegian University of Science and Technology (NTNU). He is currently PhD Research Fellow at the Department of Neuromedicine and Movement Science, NTNU, Trondheim, Norway. His research interests include deep learning-based computer vision, markerless human movement analysis, and medical technology.

Lars Adde: Lars Adde is Senior Researcher at the Department of Clinical and Molecular Medicine, NTNU, Trondheim, Norway and Physiotherapist at the Clinic of Clinical Services, Trondheim University Hospital, Trondheim, Norway. Adde holds a PhD degree in Clinical Medicine from NTNU, Master from NTNU, and Bachelor in Physiotherapy from Oslo University College (now Oslo Metropolitan University). He is Associate Director in six multisite studies predicting cerebral palsy and long-term motor outcomes by use of clinical and computerized assessment of spontaneous infant movements. His main research interest is movement analysis of young infants.

Ragnhild Støen: Ragnhild Støen is Professor at the Department of Clinical and Molecular Medicine, NTNU, Trondheim, Norway. Støen holds a PhD degree and Master M.D. from NTNU. She is Head of Neonatology at the Department of Neonatology, Trondheim University Hospital, Trondheim, Norway. Her particular interests are early neuroprotection after perinatal asphyxia, early detection of cerebral palsy in high-risk infants and ethical challenges related to infants born at the border of viability.

Heri Ramampiaro: Heri Ramampiaro is Professor at the Department of Computer Science, NTNU, Trondheim, Norway. He is Head of the Data and Artificial Intelligence (DART) research group and Head of the department. Ramampiaro holds a PhD degree in Computer Science from NTNU, M.Sc. from Stavanger University College (now University of Stavanger), and B.Eng. from Aalesund College (now NTNU). His current main research interests include machine learning, data/text mining, information retrieval, and NLP.

Espen A.F. Ihlen: Espen A. F. Ihlen is Associate Professor at the Department of Neuromedicine and Movement Science, NTNU, Trondheim, Norway. Ihlen holds a PhD degree in Clinical Medicine and M.Sc. in Human Movement Science from NTNU. His main interest is development of convolutional neural network architectures and machine learning for human movement analysis.