Computerized Medical Imaging and Graphics (2021)

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

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*

ARTICLE INFO

Article history:

Keywords:
Computer-based risk assessment
Convolutional neural networks
Infant pose estimation
Markerless motion capture
Neuromotor impairment
Video-based analysis

ABSTRACT

Assessment of spontaneous movements can predict the long-term developmental outcomes in high-risk infants. In order to develop algorithms for automated prediction of later function based on early motor repertoire, highly precise localization of segments and joints are required. Four types of convolutional neural networks were investigated on a novel infant pose dataset, covering the large variation in 1 424 videos from a clinical international community. The accuracy 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 study shows that the accuracy of the best performing infant pose estimator is similar to the inter-rater spread of human experts, while still operating efficiently. In conclusion, pose estimation of infant movements may support research initiatives on early detection of motor disorders in children with perinatal brain injuries by quantifying infant movements from video recordings with human-level accuracy.

1. Introduction

During the first months of life spontaneous infant movements may indicate later developmental difficulties, such as cerebral palsy (CP), Rett syndrome, and autism spectrum disorder (Novak et al., 2017; Einspieler et al., 2014, 2005). 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) (Ferrari 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 three 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

*Corresponding author.
e-mail: espen.ihlen@ntnu.no (Espen A. F. Ihlen)

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 impairment (e.g., 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 (Ferrari 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 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 anatomic proportions and distribution of body poses (Sciortino et al., 2017). Employed on infant images, the accuracy drops significantly, with more than 10% of the estimated body keypoint positions placed outside a head length distance from the annotated ground truth positions (i.e., 88-90% in the $PCK_h@1.0$ metric described in Section 2.3) (Sciortino et al., 2017; Hesse et al., 2018). 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) retrains the openly available OpenPose network (Cao et al., 2019) by utilizing a dataset of 9 039 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). Furthermore, recent developments in HPE outperform OpenPose by deploying novel multi-scale networks and by maintaining higher spatial resolution (Newell et al., 2016; Sun et al., 2019). Like other solutions that are targeted towards multi-person pose estimation, such as DeeperCut (Insafutdinov et al., 2016), 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 accuracy-efficiency tradeoffs 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 an accuracy 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 accuracy and computational efficiency of eight different ConvNet models, including the commonly used OpenPose network. We compare the accuracy 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 1 424 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 recordings thus follow the standards for video-based GMA (Einspieler and Prechtl, 2005). 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 20 000 video frames. The dataset emphasizes the heterogeneity in spontaneous movements by including videos from 12 different sites from seven countries across the globe (i.e., Norway, India, USA, 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, 8 000 (40%) frames originated from standardized hospital recordings, 8 000 (40%) from home-based smartphone recordings, and the remaining 4 000 (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 20 000 separate frames (8 000, 8 000, and 4 000 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 20 000 frames were split into training (14 483 (72%)), validation (1 493 (8%)), and test sets (4 024 (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 all body keypoints, through manual annotation. This resulted in a total of 380 000 human labels (i.e., 19 annotated keypoint positions for each of the 20 000 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 agreement in terms of the mean Euclidean annotation 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 $(\bar{x}_{b,i}, \bar{y}_{b,i})$, across the N (i.e., 10) experts for the S (i.e., 100) frames (see 1). H was normalized according to mean head length of an infant, defined as the distance from the top of the head to the upper neck.

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

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 state-of-the-art method for infant pose estimation, the OpenPose network (Cao et al., 2019) (see Fig. 2a for an architectural overview), was trained to yield baseline performance 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. Moreover, it employs bilinear upscaling of network outputs to improve the accuracy in the estimates of body keypoint positions. OpenPose and CIMA-Pose operate on similar image input resolutions of 368×368 pixels. 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 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., 2020b,a; Groos and Aurlien, 2018).

In the experiments, all models 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 accuracy of the models included in the comparative analysis, positions of body keypoints were predicted on the separate test set of In-Motion Poses, comprising 4 024 images. All models were evaluated using the raw model outputs, omitting the expensive multi-scale testing and flipping procedure commonly used for benchmarking HPE (Tang et al., 2018; Yang et al., 2017). Model accuracy was determined by comparing the model outputs to human annotations. The accuracy 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 ($PCK@Human^{0.95}$). $PCK_h@\tau$ computes the fraction of keypoints within τl distance from the annotated position, where l is the infant head length. To account for both model robustness and performance in high-precision pose estimation, we calculated measures of $PCK_h@\tau$ across various

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.

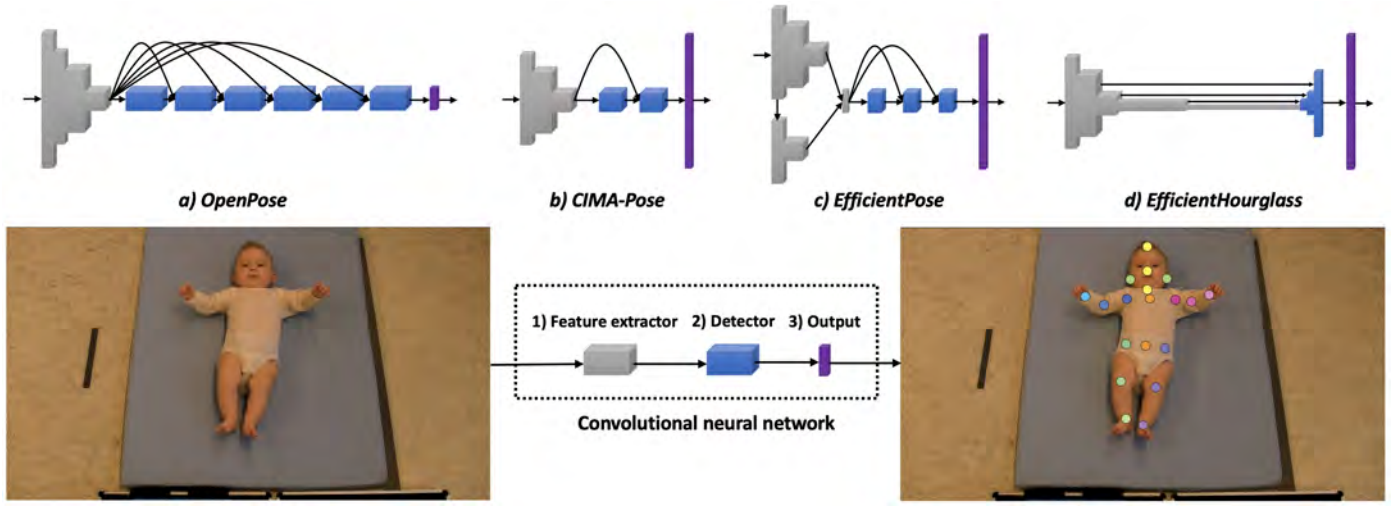

Fig. 2. ConvNets address infant pose estimation from video frames in a frame-by-frame manner by 1) extracting image features, 2) determining features relevant for detection, and 3) estimating infant keypoint positions. The height of the ConvNet blocks (i.e., feature extractor, detector, and output) indicates the block's spatial resolution in relation to the resolution of the input image (e.g., OpenPose has lower output resolution than CIMA-Pose, EfficientPose, and EfficientHourglass).

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 $PCK_h@0.1$. Moreover, the ME measure reflects the average accuracy 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 $d_{m,b,i} = \sqrt{(x_{m,b,i} - \hat{x}_{b,i})^2 + (y_{m,b,i} - \hat{y}_{b,i})^2}$ is the Euclidean distance from the estimated keypoint position $(x_{m,b,i}, y_{m,b,i})$ of model m to the human annotation $(\hat{x}_{b,i}, \hat{y}_{b,i})$, for keypoint b in image i of the test set. ME was normalized with respect to the mean head length. To compare model accuracy against human-level performance, we introduce a metric, called $PCK@Human^{0.95}$. $PCK@Human^{0.95}$ defines the percentage of model predictions within the 95th percentile of the annotation spread of human experts:

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

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

where δ is a binary step function with ${H_b}^{0.95}$, which should

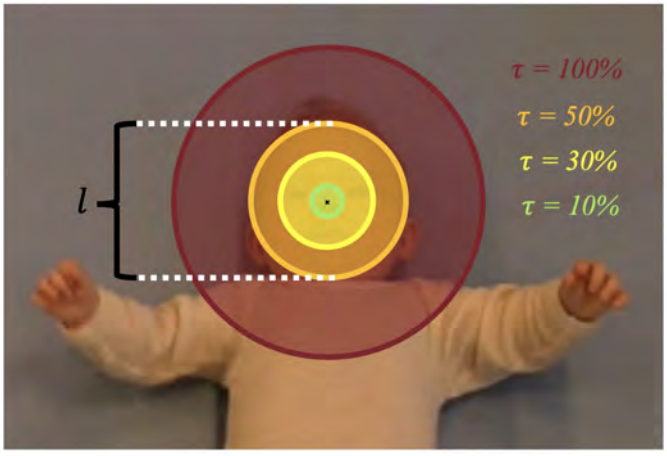

Fig. 3. $PCK_h@\tau$, **the percentage of predictions within** τl **distance from the ground truth location (e.g., nose), is computed across four different thresholds** τ **(i.e., 100%, 50%, 30%, and 10%), evaluating the accuracy of a model, from coarse to fine.**

not be mistaken for the mean annotation spread H_b, defining the 95th percentile of the Euclidean distance of the annotations of an individual human expert to the average human annotations. Thus, $PCK@Human^{0.95} = 95\%$ reflects human-level accuracy. 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 human annotation 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 accuracy.

In addition to model accuracy, 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. We used a batch size of 10 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[1] to 100 images to the full training set of 14 483 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 accuracy in ME, $PCK_h@1.0$, $PCK_h@0.5$, $PCK_h@0.3$, 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. The sample 11.6K* comprised the 80% randomly selected frames in the full training set.

[1]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.

3. Results

Table 1 displays the performance of the eight different ConvNets included in the comparative analysis. In terms of accuracy, 30-52% decrease in ME compared to the OpenPose baseline is achieved. This is supported by 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 58.87% to 81.20% are displayed, compared to 34.73% for OpenPose. With regards to computational efficiency, all models are smaller, with 1.4-54 times fewer parameters, and require less computation than OpenPose, i.e., 2.2-168 times less FLOPs. Moreover, for the most computationally efficient ConvNet, EfficientPose RT, 16 times speed-up of inference is achieved.

Table 2 displays the accuracy of the top-performing ConvNet of each model family. The most accurate model, EfficientHourglass B4, achieved an ME of 0.0654 compared to the average human annotation spread H of 0.0525. This equals an average percentage of human-level performance (i.e., $PCK@Human^{0.95}$) of 87.02%, compared to 43.07% 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 all ConvNets benefit from increased training set size. However, whereas accuracy of OpenPose and CIMA-Pose saturates at sample sizes beyond 5 000 images, EfficientPose III and EfficientHourglass B4 benefit from larger training sets. Moreover, there is a tendency that EfficientPose III and EfficientHourglass are more stable across dataset sizes, with a smaller difference in accuracy from 100 to 14 483 images, compared to OpenPose and CIMA-Pose.

4. Discussion

The main objective of the study was to obtain computationally efficient markerless infant pose estimation with a level of accuracy approaching that of human expert annotations. A comparative analysis has showed that accuracy levels comparable to human expert performance can be achieved, by utilizing contemporary ConvNets for HPE together with an extensive infant video database. This is reflected by $PCK@Human^{0.95}$ of

Table 1. The performance of the different ConvNets, pretrained on MPII (Andriluka et al., 2014) and fine-tuned on In-Motion Poses, in relation to OpenPose, in terms of accuracy on the test set of In-Motion Poses, and computational efficiency from run-time experiments on an NVIDIA GTX 1080 Ti GPU

		Accuracy					Computational efficiency		
Model	**Resolution**	**@1.0**[a]	**@0.5**[a]	**@0.3**[a]	**@0.1**[a]	**ME**	**Parameters**	**FLOPs**	**Latency**
OpenPose	368×368	99.90%	99.36%	94.56%	34.73%	0.1372	25 994 358	160 588 886 296	760.97 ms
CIMA-Pose	368×368	99.98%	99.83%	98.60%	58.87%	0.0958	2 380 495	15 645 092 494	200.41 ms
EfficientPose RT	224×224	99.96%	99.69%	98.15%	58.63%	0.0966	481 336	955 490 248	47.47 ms
EfficientPose I	256×256	99.98%	99.82%	98.76%	60.03%	0.0936	743 476	1 785 432 722	80.19 ms
EfficientPose II	368×368	99.98%	99.89%	98.50%	61.91%	0.0919	1 759 372	7 944 292 598	264.84 ms
EfficientPose III	480×480	99.99%	99.94%	99.56%	78.14%	0.0705	3 258 888	23 777 830 318	683.52 ms
EfficientPose IV	600×600	99.98%	99.92%	99.44%	71.06%	0.0798	6 595 430	73 621 311 041	1824.62 ms
EfficientHourglass B4	608×608	99.99%	99.95%	99.56%	81.20%	0.0654	18 699 936	27 009 544 472	696.99 ms

[a] PCK_h@1.0, PCK_h@0.5, PCK_h@0.3, and PCK_h@0.1 are abbreviated as @1.0, @0.5, @0.3, and @0.1, respectively.

Table 2. The accuracy 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 accuracy (i.e., inter-rater spread H) across body parts b, as evaluated by the proposed $PCK@Human^{0.95}$ metric.

			PCK@Human$^{0.95}$			
b	**H_b**	**$H_b^{0.95}$**	**OpenPose**	**CIMA-Pose**	**EfficientPose III**	**EfficientHourglass B4**
Head top	0.0546	0.1214	44.14%	58.30%	85.41%	90.58%
Nose	0.0277	0.0525	13.49%	37.20%	69.09%	80.64%
Right ear	0.0544	0.1450	61.51%	84.42%	91.58%	89.21%
Left ear	0.0489	0.1404	56.81%	78.16%	91.03%	91.10%
Upper neck	0.0521	0.1171	54.27%	83.23%	88.07%	89.96%
Right shoulder	0.0527	0.1121	45.05%	75.42%	87.52%	86.90%
Right elbow	0.0425	0.0952	33.97%	70.83%	84.22%	88.20%
Right wrist	0.0369	0.0819	23.71%	58.77%	80.02%	82.58%
Upper chest	0.0645	0.1230	52.88%	76.54%	82.58%	83.55%
Left shoulder	0.0563	0.1160	43.09%	62.38%	87.30%	88.12%
Left elbow	0.0413	0.0948	31.34%	48.96%	82.11%	85.54%
Left wrist	0.0365	0.0789	22.29%	46.07%	71.69%	80.09%
Mid pelvis	0.0781	0.1642	70.68%	84.57%	88.39%	90.90%
Right pelvis	0.0813	0.1616	66.18%	84.34%	88.25%	89.76%
Right knee	0.0556	0.1129	45.00%	77.49%	86.63%	88.79%
Right ankle	0.0408	0.0838	26.24%	58.47%	72.61%	79.67%
Left pelvis	0.0836	0.1647	65.28%	81.44%	89.21%	91.80%
Left knee	0.0489	0.1051	34.97%	50.60%	88.62%	89.91%
Left ankle	0.0404	0.0880	27.49%	48.19%	78.43%	86.11%
All body parts	*0.0525*	*0.1136*	*43.07%*	*66.60%*	*83.83%*	*87.02%*

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 [-0.01, 0.03]	0.08 [-0.04, 0.34]	0.52 [-0.03, 0.86]	0.69 [0.18, 0.97]
$ICC(C,1)$	-0.01 [-0.46, 0.43]	0.47 [0.03, 0.76]	0.94 [0.85, 0.98]	0.96 [0.89, 0.98]

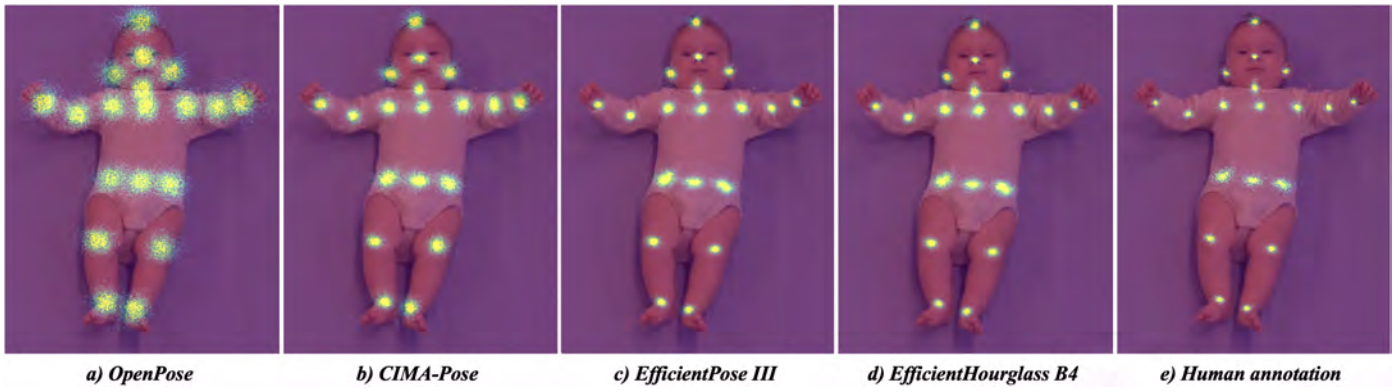

Fig. 4. From left: a-d) The distribution of model prediction errors of the different ConvNets on the test set of In-Motion Poses across body parts, and e) the distribution of the annotation spread of the 10 human experts across 100 inter-rater frames.

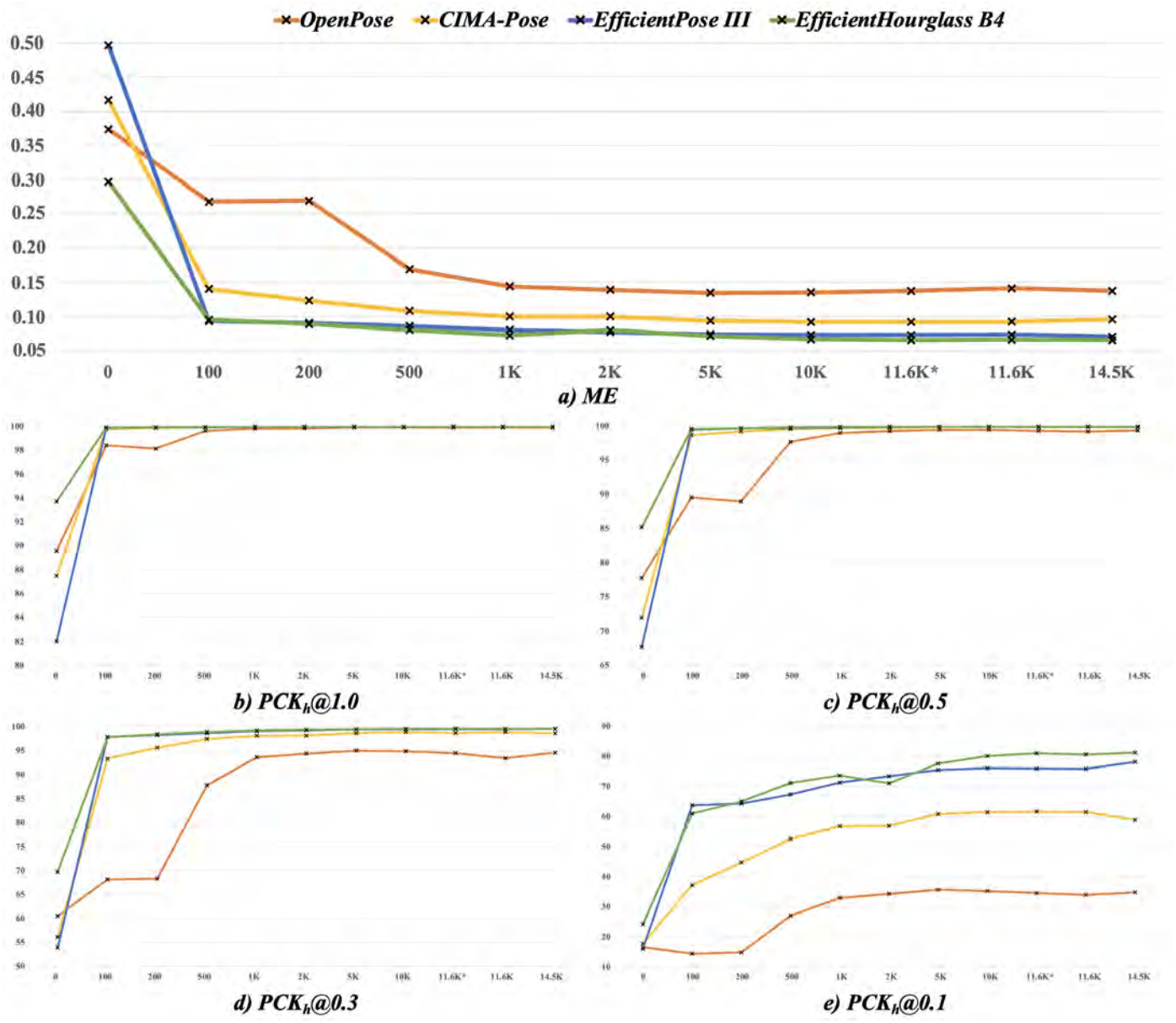

Fig. 5. Accuracy of OpenPose, CIMA-Pose, EfficientPose III, and EfficientHourglass B4, all pretrained on MPII (Andriluka et al., 2014), with increasing amounts of data for fine-tuning on In-Motion Poses.

the top-performing ConvNets approaching human-level accuracy, whereas the commonly applied OpenPose network does not reach similar level of accuracy.

4.1. Improving accuracy

The large improvement in accuracy compared to existing studies (Sciortino et al., 2017; Hesse et al., 2018; Chambers et al., 2020) 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 accuracy of OpenPose from 90% to 99.90% and from 76% to 99.36%, on PCK_h@1.0 and PCK_h@0.5, respectively (Sciortino et al., 2017), as illustrated by Fig. 5. Taking into account the error taxonomy of Ruggero Ronchi and Perona (2017), this indicates that the coarse localization errors have been reduced, in particular the frequency of inversions (i.e., the predictions that appear at a wrong body keypoint, such as misinterpretation of the left and right wrist) or misses (i.e., the erroneous localizations that are made without interfering with other keypoints). Despite the increased robustness with regards to coarse prediction errors, the optimal level of accuracy 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. 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@0.1 and $PCK@Human^{0.95}$ of CIMA-Pose, EfficientPose, and EfficientHourglass B4, compared to OpenPose, is due to a reduction in jitter. Jitter is the most prominent type of error in ConvNets for pose estimation (Ruggero Ronchi and Perona, 2017), reflecting position estimates that are slightly inaccurate related to ground truth positions. CIMA-Pose, EfficientPose[2], and EfficientHourglass B4 reduce prediction jitter 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 $PCK@Human^{0.95}$. 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, post-processing of the frame-by-frame position estimates over consecutive frames by low-pass filters, such as median filtering (Tukey, 1977), might reduce jitter more effectively. However, this demands that the video has a sufficient sample rate (e.g., 60 fps). Furthermore, jitter may also be minimized by improving the accuracy of annotated keypoint positions. As illustrated in Fig. 4, the distributions of prediction errors of EfficientPose III and EfficientHourglass B4 across body parts resemble the annotation 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 jitter, 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 accuracy. 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 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 accuracy improvements beyond jitter, 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 accuracy against metrics related to human-level performance, such as $PCK@Human^{0.95}$, to evaluate the progress on these tasks in relation to human-level accuracy.

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 (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 3.19% decrease in $PCK@Human^{0.95}$, compared to EfficientHourglass B4, despite having 5.7 times fewer parameters. Combining this observation with the influence of high input and output resolution on accuracy, we would suggest further studies to investigate the effect of high resolution with low-complexity ConvNets. This could potentially narrow the current gap in accuracy 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 accuracy and computational efficiency, by carefully assessing the accuracy of

[2]EfficientPose IV displayed lower accuracy than EfficientPose III, which we suspect was due to small batch size (i.e., 5) during training due to limitations in GPU memory, causing poor convergence.

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 (Ferrari 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 above 20 fps (i.e., 47 ms latency of EfficientPose RT, as shown by Table 1) on an NVIDIA GTX 1080 Ti consumer GPU. Taking into account that the authors of OpenPose report 36 ms latency of OpenPose on a similar GPU (Cao et al., 2019), whereas we achieve 761 ms (more than 20 times slower) with an unaccelerated TensorFlow model, there are potentials for significant speedups of the pool of models studied in this paper. This may be obtained by implementing the ConvNets in low-level code like C++ or CUDA. Thus, a three-minute video of spontaneous infant movements could potentially be processed 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 accuracy. Quantization-aware training, knowledge distillation, model pruning, and sparse kernels are paths that are worth to investigate (TensorFlow, 2020; Buciluǎ 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 most accurate ConvNet on the dataset proposed by Hesse et al. (2018), the general performance of infant pose estimation was acceptable, only with few minor deviations in the localization of wrists (Fig. B.7). However, further studies should more thoroughly assess the external validity of the infant pose estimator to verify that the high accuracy-level demonstrated by the present quantitative analysis indeed can be reproduced. This involves assessing the robustness of the infant pose estimators 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 accuracy 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 mentioning 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 accuracy. 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 three 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. High-precision 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 motor disabilities. With no dependency to body-worn markers, sensors or other expensive laboratory equipment, the automatic infant pose estimation can

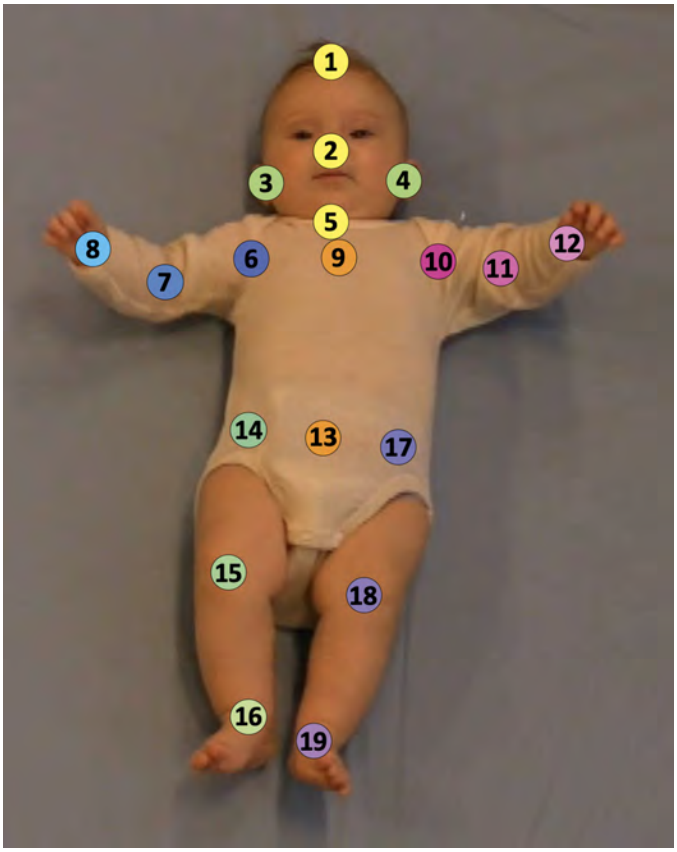

Fig. A.6. The placements of the 19 different body keypoints on an infant.

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.

Appendix A. Keypoint definitions

The set of 19 body keypoints along with their definitions (see Fig. A.6 and Table A.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. All body keypoints were annotated in all images regardless of their type of visibility (i.e., visible or occluded).

Appendix B. External validation

To investigate the external validity of the most accurate ConvNet in our comparative analysis, namely EfficientHourglass B4, we conducted a qualitative experiment on an openly available dataset of 12 synthetic infant video recordings proposed by Hesse et al. (2018). More specifically, the EfficientHourglass B4 model, fine-tuned on the training set of In-Motion Poses, estimated the locations of the 19 body keypoints in these videos, which are of quite different nature than the recordings in In-Motion Poses (Fig. 1). The model predictions on a randomly selected frame in each of the 12 infant videos are provided in Fig. B.7.

Table A.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

Acknowledgments

This study was possible only due to the unified In-Motion research initiative on computer-based assessment of infant spontaneous movements and prediction of cerebral palsy, resulting in the multi-site database of infant recordings. The authors would like to acknowledge the following key personnel and institutions contributing in collecting video recordings: Norway; Toril Larsson Fjørtoft at St. Olavs University Hospital, Inger Elisabeth Silberg at Oslo University Hospital, Nils Thomas Songstad at University Hospital of North Norway, Angelique Tiarks at Levanger Hospital, Henriette Paulsen at Vestfold Hospital Trust, India; Niranjan Thomas at Christian Medical College Vellore, United States; Colleen Peyton at University of Chicago Comer Children's Hospital, Raye-Ann de Regnier and Lynn Boswell at Ann & Robert H Lurie Children's Hospital of Chicago, Turkey; Akmer Mutlu at Hacettepe University, Belgium; Aurelie Pascal at Ghent University, Denmark; Annemette Brown at Nordsjællands Hospital Hillerød, United Kingdom; Anna Basu at Newcastle upon Tyne Hospitals. This work was supported by the Liaison Committee between the Central Norway Regional Health Authority and the Norwegian University of Science and Technology (SO: 90056100), the Joint Research Committee between St. Olavs University Hospital and the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, and RSO funds from the Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology.

Fig. B.7. Predictions of keypoint locations in randomly selected frames from videos in the MINI-RGBD dataset (Hesse et al., 2018).

References

Adde, L., Brown, A., Van den Broeck, C., De Coen, K., Horsberg Eriksen, B., Fjørtoft, T., Groos, D., Ihlen, E., Osland, S., Pascal, A., et al., 2021. The in-motion-app for remote general movement assessment: A multi-site observational study. BMJ Open, In press .

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. Developmental Medicine & Child Neurology 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, in: Advances in Neural Information Processing Systems, pp. 3027–3038.

Buciluǎ, 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 Transactions on Pattern Analysis and Machine Intelligence .

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 au-

tomatically assess infant neuromotor risk. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, 2431–2442.

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

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

Einspieler, C., Peharz, R., Marschik, P.B., 2016. Fidgety movements–tiny in appearance, but huge in impact. Jornal de Pediatria 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. Mental retardation and developmental disabilities research reviews 11, 61–67.

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. Research in Autism Spectrum Disorders 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., Einspieler, C., Prechtl, H.F., BOS, A., Cioni, G., 2004. Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants. Mac Keith Press.

Ferrari, F., Plessi, C., Lucaccioni, L., Bertoncelli, N., Bedetti, L., Ori, L., Berardi, A., Della Casa, 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. Journal of clinical medicine 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. Physical therapy 99, 701–710.

Groos, D., Adde, L., Ihlen, E., 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. Applied Intelligence .

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), pp. 0–0.

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. Journal of Clinical Medicine 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 statewide geographical study. Journal of paediatrics and child health 55, 548–554.

Morgan, C., Romeo, D.M., Chorna, O., Novak, I., Galea, C., Del Secco, 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. Journal of clinical medicine 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 pediatrics 171, 897–907.

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, in: 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 on: 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. Computer Vision and Image Understanding 140, 30–42.

Ruggero Ronchi, M., Perona, P., 2017. Benchmarking and error diagnosis in multi-instance pose estimation, in: Proceedings of the IEEE international conference on computer vision, pp. 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., 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. Journal of clinical medicine 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. Pediatric Research 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 on: 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. volume 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 on: 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.