A 3D pose estimation framework for preterm infants hospitalized in the Neonatal Unit

Ameur Soualmi, Christophe Ducottet, Hugues Patural, Antoine Giraud, Olivier Alata

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16333-6

A 3D pose estimation framework for preterm infants hospitalized in the Neonatal Unit

Ameur Soualmi[1,2] · Christophe Ducottet[1] · Hugues Patural[2,3] · Antoine Giraud[2,3] · Olivier Alata[1]

Received: 21 July 2022 / Revised: 13 June 2023 / Accepted: 17 July 2023

Abstract
Infant pose estimation is crucial in different clinical applications, including preterm automatic general movements assessment. Recent infant pose estimation methods are limited by a lack of real clinical data and are mainly focused on 2D detection. We introduce a stereoscopic system for infants' 3D pose estimation, based on fine-tuning state-of-the-art 2D human pose estimation networks on a large, real, and manually annotated dataset of infants' images. Our dataset contains over 88k images, collected from 175 videos from 53 premature infants born <33 weeks of gestational age (GA), acquired within the Neonatology department of the Centre Hospitalier Universitaire de Saint Etienne, France, between 32 and 41 weeks of GA. This framework significantly reduced the pose estimation error compared to existing 2D infant pose estimation networks. It achieved a mean error of 1.72 cm on 18000 stereoscopic images in the 3D pose estimation task. This framework is the first 3D pose estimation tool dedicated to preterm infants hospitalized in the Neonatal Unit that does not depend on any visual markers or infrared cameras.

Keywords Infant pose estimation · 3D pose · Stereoscopic vision · General movements assessment · Premature infants · Preterm birth · Video analysis

✉ Ameur Soualmi
ameur.soualmi@univ-st-etienne.fr

Christophe Ducottet
ducottet@univ-st-etienne.fr

Hugues Patural
hugues.patural@chu-st-etienne.fr

Antoine Giraud
antoine.giraud@univ-st-etienne.fr

Olivier Alata
olivier.alata@univ-st-etienne.fr

1 Laboratoire Hubert Curien, CNRS UMR 5516, Université Jean Monnet, IOGS, Saint-Étienne, France

2 INSERM, U1059 SAINBIOSE, Université Jean Monnet, Saint-Étienne, France

3 Service de Néonatalogie, Centre Hospitalier Universitaire de Saint-Étienne, Saint-Étienne, France

Published online: 14 August 2023

1 Introduction

Preterm birth, occurring before 37 weeks of gestational age, is a risk factor for developmental disabilities, such as behavioral difficulties, cognitive impairments, or cerebral palsy (CP) [29]. In many cases, these developmental disabilities cannot be identified before two years of age [29]. However, early identification of preterm infants with abnormal developmental trajectories is critical to initiate early developmental intervention, to prevent the occurrence of such developmental disabilities [36].

The General Movement Assessment (GMA) analyses the complexity, variability, and fluidity of preterm spontaneous movements using video recordings [9]. The GMA is a reliable assessment of brain maturation and is able to identify preterm infants with abnormal developmental trajectories [26]. However, this method remains subjective and time-consuming. This motivates the need for automated GMA, especially with the continuous progress in computer vision and artificial intelligence tools. To achieve this goal, a primary task is to track preterm infant movements and estimate their poses in the three dimensions of space using a markerless tool adapted to this specific population.

Thanks to the advances in deep learning, many reliable neural models are available for automatic 2D and 3D human pose estimation [6, 8, 38]. However, as these models are trained on general-purpose datasets containing mainly adult persons [2, 20], they perform poorly for infants who are anatomically different. To solve this problem, researchers retrained 2D neural networks either on synthetic datasets [16] or on datasets collected from the internet [7]. In both cases, the training data was not representative enough of the target population of preterm infants hospitalized in the Neonatal Unit. Moreover, 2D movement analysis remains limited since it does not exploit the overall infants' movement information in space, and information loss can go up to 53% due to dimensionality reduction [3]. On the other hand, 3D analysis can be advantageous since it provides a complete analysis of infant movements, which allows an objective classification of these movements as normal and abnormal according to their complexity, variability, and fluidity. The purpose of this work was first to introduce a real and representative dataset of preterm infants and second to propose a markerless, accurate, and safe 3D pose estimation framework to automatically assess preterm infants' spontaneous movements.

The main contributions of this work are (1) to propose the first framework for 3D infant pose estimation from stereoscopic images dedicated to preterm infants hospitalized in the Neonatal Unit; (2) to create a fully annotated dataset of real preterm infants' images with a clinical protocol that follows the guidelines for the GMA [9]; (3) to provide models for the main state of the art convolutional neural network (CNN) architectures specifically retrained with the new dataset;[1] (4) to perform an evaluation of these models for 2D and 3D infant pose estimation, providing a new baseline in this field.

2 Related work

2.1 2D pose estimation for infants

During the last two decades, a tremendous effort was made toward developing automatic tools to assess the general movements of preterm infants. In the beginning, many researchers

[1] The retrained networks can be downloaded from here: https://drive.google.com/drive/folders/1SEuTqrNdz6ubRGwMaUazil0BVOqf2cYw?usp=sharing

tested motion sensors, including accelerometers [12, 13, 25] or electromagnetic tracking systems (EMTS) [17, 18] to track infants' movements.

The advances in computer vision techniques encouraged other groups to use visual sensors. Some researchers opted for marker-based approaches [4, 23], which can be accurate but need a specific setup that may not be practical. On the other hand, markerless approaches became the state-of-the-art methods for infant pose estimation since they are cost-effective and due to the continuous improvement of computer vision techniques.

Adde et al [1] used motion images which were calculated as the difference between two consecutive frames where each pixel represents a point value of 0 and 1, 0 being black and representing no movement, and 1 being white and representing movement [1]. These motion images were used to calculate Motiongrams, Motion quantities, and motion centroids. A step further was made with the use of optical flow algorithms. Stahl et al [37] performed motion extraction using optical flow then feature extraction using wavelets and frequency analysis to classify infants' fidgety movements. Many others opted for large displacement optical flow (LDOF) [27, 30, 31] to track the movements and obtain the movements' velocities. These methods have many limitations, especially in the case of occlusions and illumination changes.

In the last decade, convolutional neural networks have revolutionized the task of pose estimation for adults and achieved very promising results on the different state-of-the-art datasets (e.g. COCO [20] or MPII [2]). This inspired many works on infants' pose estimation [11, 22, 24, 32, 33], and all of them are based on the Openpose architecture [6] which was trained on datasets containing only images of adults who are structurally and anatomically different compared to infants [34].

To address this problem, Chambers et al. [7] have retrained the Openpose network with their own labeled dataset of infant pose consisting of 9039 infant images collected partially from videos on Youtube and from clinical data, which reduced the mean error by 60%. In addition to the fact that Openpose lacks sufficient scaling of network depth and its computational insufficiency [14], the ground truth body joints annotation process using Vatic[7] was not very accurate and was missing many occluded and not visible joints as can be seen from their publicly available dataset.

The absence of open-source infants pose images mainly for privacy concerns is a problem that many researchers have tried to address. Hesse et al. [15] introduced a synthetic dataset of infants, called MINI-RGBD for Moving INfants In RGB-D, created from only 12 RGB-D videos of moving infants using a textural mapping to a 3D model mesh called SMIL for Skinned Multi-Infant Linear model. Both its simplicity and exclusively synthetic nature cause the pose estimation models trained on MINI-RGBD not to generalize well on real-world infant images [16]. This is why Huang et al. [16] created a hybrid synthetic and real infant pose dataset (SyRIP) from real images collected from YouTube and Google images. They used a similar approach to SMIL, by fitting the 3D model mesh to a real infant image and getting synthetic images with changing backgrounds, texture maps, lightning, and camera position, resulting in 700 real and 1000 infant synthetic images. This small dataset was used to train an architecture (FIDIP) that contains a domain classifier that promotes a feature extractor to retain the ability to extract keypoints information but also to ignore the differences between the real and the synthetic input images [16].

The group tested the effect of retraining three existing models as backbone networks on their SyRIP dataset in different ways, but what we found questionable was mainly the small number of images used for testing (100 & 500) and the use of mean average precision (mAP) evaluation metric with OKS thresholds [20] which was previously implemented specifically for the COCO dataset (explained in Section 3.2). In this study, we provide a PCK-based

evaluation of this model using our dataset containing real images captured in a clinical environment.

2.2 3D pose estimation for infants

For an accurate infant movement assessment and to achieve a complete analysis of infant movements in space, 3D infant pose estimation is an essential task.

Meinecke et al. [23] were the first to address this approach using an analysis system Vicon 370 motion, consisting of 7 infrared cameras on tripods and many reflective markers placed on different body joints. Even though a very good spatial precision of 2 mm was achieved, the system still has a difficult setup that is impractical for clinical use.

Other researchers tried to obtain a 3D keypoint representation using RGB images in addition to depth images obtained from depth cameras. Wu et al. [39] used Kinect to capture color and depth videos of infant movements. Then used OpenPose [6] to estimate infant 2D movement from RGB images. The 3D coordinates of the infant's joints are achieved after combination with corresponding depth images. Li et al. [19] used exactly the same approach with the same pose estimation network and image acquisition protocol but with a correction of depth information to solve the problem of matching between RGB and depth images.

Even if these approaches are simpler and faster than multi-camera motion capturing systems, however when dealing with complex infant pose with a harder view or obstructive gesture, the method may be misidentified [40]; depth values captured by Kinect represent the body surface information [39] and the depth image will map a joint X and Y to the same depth value if one joint is occluded by another. Also, the use of infrared light-emitting cameras may have a health impact on infants [21].

The only work to our knowledge that presented a stereoscopic image acquisition protocol for infant pose estimation was presented by Shivakumar et al. [35]. They used two stereoscopic cameras, one placed right above the baby and the other positioned on the baby's right side. To obtain depth images, the team used a series of calibration, rectification, and matching operations to get disparity maps and then depth maps.

Since the group used only a tracking algorithm, infants were provided with a blue onesie to facilitate the tracking of the torso center using the segmented blue color mask. The optical flow was used for tracking limbs, and a manual selection of regions was needed where the user clicked on a point within the limb region; then, a marker was set using their marker identification method.

3 Proposed 3D pose estimation framework

We propose a 3D pose estimation framework based on stereoscopic imaging (see Fig. 1). Stereoscopic video sequences of the infants were recorded through a stereoscopic camera located above the infant. Each pair of frames were subsequently rectified and processed by a 2D pose estimation model, giving 17 keypoints per image located at joint positions. A stereo triangulation was then performed to recover the 3D position of the joints. The output of the system is thus a temporal sequence of 3D joint positions.

The ZED 2 stereoscopic camera (Stereolabs, San Francisco, CA)system was chosen due to its simplicity, reliability, and the fact that it does not affect the infant's health. The originality of our framework is to perform a joint-based 3D reconstruction. Instead of reconstructing the depth of all the points of the scene using a stereo association algorithm, we only recon-

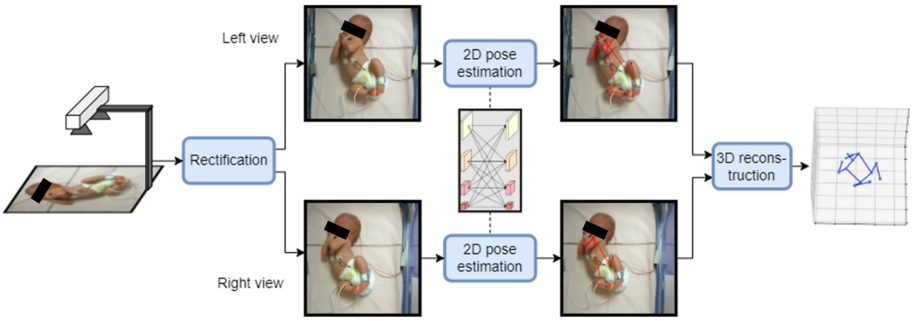

Fig. 1 3D infant pose estimation framework based on stereoscopic imaging, 2D pose estimation and 3D reconstruction using triangulation

structed the depth of the body joints obtained separately on the two rectified views by direct triangulation [20]. This reconstruction is indeed much more accurate since it can infer the true depth of the joints even if the latter are not directly visible on images. Comparatively, classical association-based algorithms reconstructed the depth of the points located on the surface of the body [39] and were not able to recover the true joint positions. More precisely, if the stereo system has been calibrated, the depth is recovered by direct triangulation (Fig. 2):

$$Z = \frac{Bf}{D} \tag{1}$$

where, B and f are the camera baseline and focal length respectively, and D is the point disparity.

This framework strongly relies on the precise estimation of the 2D location of the body joints, which depends on two key components: (1) the dataset used for training the model and (2) the deep learning model itself.

3.1 Dataset

The AGMA dataset was built by recording real images of 53 premature infants enrolled in the AGMA[2] study, born from October 2020 to June 2021 (see Fig. 3). All infants born before 33 weeks of gestational age (GA) and hospitalized in the Neonatology department of the Centre Hospitalier Universitaire de Saint-Étienne, France, were included. Exclusion criteria were the ongoing presence of ventilatory support, contraindication of a radiant heat warmer, and absence of written parental consent. The features of the included children are summarized in Table 1.

The video recording protocol was controlled according to Prechtl's method of GMA [9]. Infants wearing a diaper were placed in a radiant heat warmer in a supine position. The room light illuminance was controlled between 60 and 120 lux. Videos were recorded for one hour using the ZED 2 (Stereolabs, San Francisco, CA) stereoscopic camera positioned perpendicularly to the warmer, with frames of 1280x720 resolution at 30 FPS.

Each included infant was recorded from one to three times, with a period $\geq$ 7 days between two consecutive recordings. The dataset contains 44,250 stereoscopic images (88,500 in total) from a selection of 175 videos from 53 infants acquired between 32 and 41 weeks of GA. All

[2] Élaboration d'un outil pronostique développemental de l'enfant prématuré basé sur l'analyse automatisée de la motricité spontanée; IDRCB 2020-A03335-34.

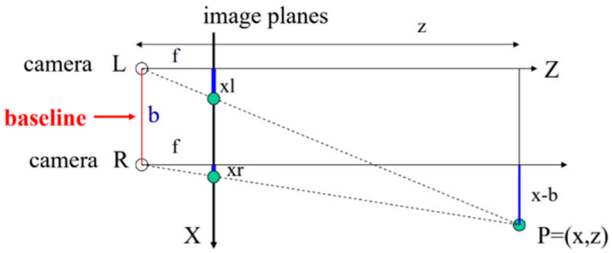

Fig. 2 Stereo vision diagram using an aligned pair of cameras

images were manually annotated and reviewed. The keypoints annotation process was done according to the COCO keypoint detection task, using 17 points forming the infant skeleton with a bounding box around them [20]. The annotation process was performed manually by a group of annotators specifically trained for this task. They were given a composite video containing the two rectified left and right views, and the task consisted of successively pointing each skeleton joint on both views for each video frame. The consistency of the annotation across time was subsequently checked for each video. Finally, the 3D annotation was obtained by triangulation using (1). AGMA dataset was divided into three independent sets for training, validation, and testing. Each set contains distinct subjects with various poses, skin colors, and ages. The training set is composed of 156 stereoscopic videos of 5 seconds duration (46800 frame pairs), the validation set is composed of 13 videos of 5 seconds duration (4500 frame pairs), and the test set comprises of 10 videos of one minute (36000 frame pairs).

It should be noted that the AGMA dataset cannot be directly compared to state-of-the-art publicly available datasets such as SyRIP and MINI-RGBD as our objective was to capture images following Prechtl's method within a clinical setting. The MINI-RGBD dataset comprises images derived from 12 entirely synthetic videos featuring simplistic infant poses

Fig. 3 A snippet of AGMA dataset images collected in a clinical environment

Table 1 Dataset population features

Male sex, n (%)	31 (58%)
Birth Gestational age, weeks, mean (SD)	30 (0.3)
Birth weight, g, mean (SD)	1357 (40)
Birth weight z-score, mean (SD)	-0.23 (1.02)
Birth length, cm, mean (SD)	38.6 (3.9)
Birth length z-score, mean (SD)	-0.48 (1.01)
Head circumference, cm, mean (SD)	27.7 (2.4)
Head circumference z-score, mean (SD)	0.01 (1.07)

Abbreviation: SD, standard deviation

[16], whereas SyRIP is a partially synthetic dataset that does not accurately represent the context and poses encountered in the GMA assessment method (see Fig. 4). Thus, the AGMA dataset stands as the first dataset captured within a genuine clinical environment, adhering to a controlled protocol that aligns with Prechtl's method of GMA [9]. It specifically concerns infants population aged between 32 and 41 weeks of gestational age which is very hard to record due to medical constraints and practical considerations which makes the AGMA unique and the first of its kind. Table 2 summarizes the characteristics of the three datasets.

3.1.1 Ethics

The AGMA study (IDRCB 2020-A03335-34) was approved by the Comité de Protection des Personnes - Sud-Est II Ethical Committee in February 2021. Written parental consent was obtained from each participant. The study was conducted in accordance with international ethical standards and the Declaration of Helsinki.

3.2 2D pose estimation

For the 2D pose estimation task, three different deep neural architectures were studied and evaluated, originally created for adult human pose estimation based on the High-Resolution Network HRNet [38]: HRNet itself, HigherHRNet, and DarkPose. These networks were chosen since they are still top ranked on different human pose estimation challenges, and they use parallel networks of different resolutions instead of traditional in-series high-to-low networks.

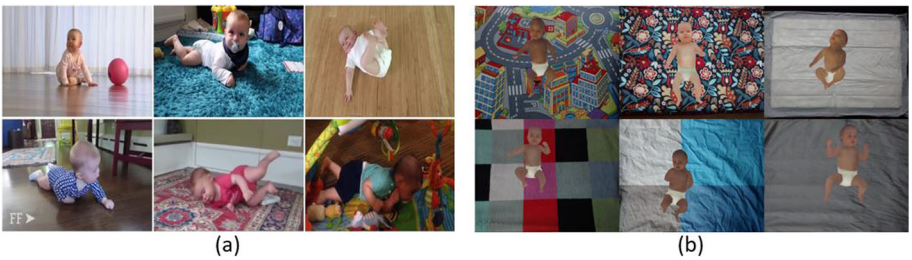

Fig. 4 (a): A snippet of SyRIP real images, (b): A snippet of MINI-RGBD images

Table 2 Comparison of AGMA dataset with other publicly available datasets

Dataset	N of images	Type	Source	Context	Age	Condition
MINI-RGBD	12000	Synthetic	SMIL [15]	Children's hospital	0-6 months	Constraint conditions
SyRIP	1700	Synthetic & real	Youtube & Google	Extra hospital	Not available	In the wild
AGMA	88500	Real	Clinical	Neonatology department	32-41 weeks of GA	GMA Protocol

Abbreviations: GA, gestational age; GMA, general movement assessment; n, number
MINI-RGBD [15] stands for Moving INfants In RGB-D dataset and SyRIP [16] for Synthetic and Real Infant Pose dataset

3.2.1 HRNet

HRNet was introduced by Ke Sun et al. [38] for different tasks, including human pose estimation, segmentation, and object detection. It became a state-of-the-art top-down pose estimation network due to the strategy of using parallel networks of different resolutions instead of traditional in-series high-to-low networks (Fig. 5). The network calculates the high resolution sub-network in parallel with lower resolution sub-networks. Then, the sub-networks are fused through the fuse layers such that each of the high-to-low resolution representations receives information from other parallel representations over and over, leading to rich high resolution representations. Maintaining high-resolution representations through the entire network makes the architecture suitable for infant pose estimation.

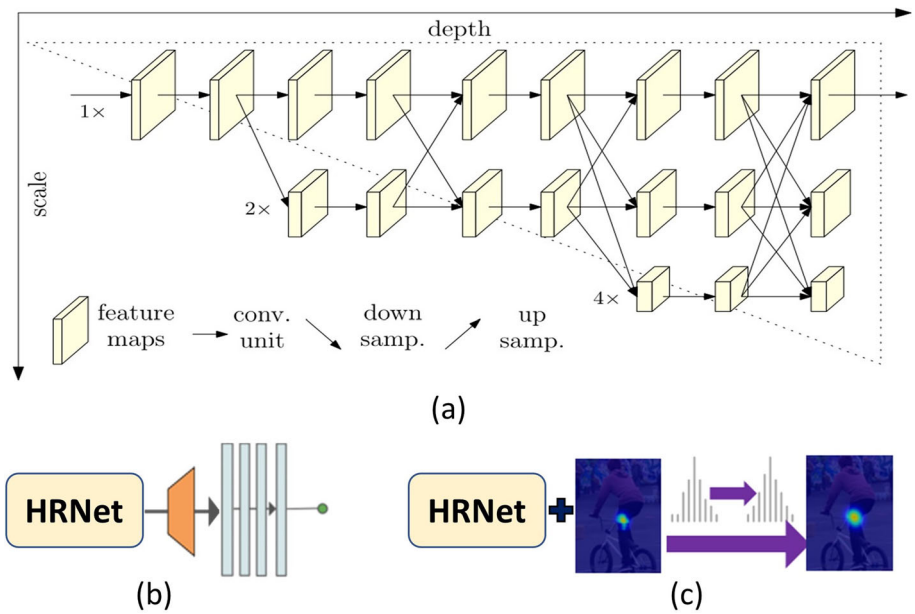

Fig. 5 2D pose estimation networks architectures: a) HRNet [38], b) HigherHRNet [8] (HRNet + deconvolution module) c) DarkPose [41] (HRNet + heatmap modulation and maximum re-localization)

For this architecture, Ke Sun et al. [38] have created four different networks, HRNet32 and HRNet48 with an input size of 256 × 192 and 384 × 288 for each. These networks were initialized by the weights of the models pre-trained on the ImageNet dataset and then trained on the COCO train2017 dataset, including 57K images and 150K person instances.

3.2.2 HigherHRNet

HigherHRNet [8] is built on HRNet as a backbone by adding a deconvolution module to predict heatmaps at multiple and higher resolutions (Fig. 5). It uses bilinear interpolation to upsample all the predicted heatmaps with different resolutions to the resolution of the input image and averages the heatmaps from all scales for the final prediction which can be beneficial for subjects of small scales in the image.

3.2.3 DarkPose

The added value in this architecture is the fact that it can be used as a plugin that improves the performance of SOTA Human pose estimation models. Taking into consideration the relevance of the output heatmap decoding process, it improves the AP of HRNet architecture by 3.8 % at least on the COCO validation set [41]. This is achieved by a series of heatmap modulations since often heatmaps predicted by a human pose estimation model do not have a good Gaussian structure compared to the training heatmap data. Then, instead of using the standard coordinate decoding method as in (2) to predict joints location, they perform a maximum re-localization with a Taylor expansion to find the maximal activation of the modulated heatmap as in (3) and finish with resolution recovery operations (Fig. 5).

$$p = m + 0.25 \frac{s - m}{\|s - m\|_2} \tag{2}$$

where m and s are the coordinates of the maximal and second maximal predicted heatmap activations

$$\mu = m - (D''(m))^{-1} D'(m) \tag{3}$$

where D'' and D' are the derivatives of the modulated heatmap, and m is the coordinates of the maximal activation of the predicted heatmap

3.3 Evaluation metrics

Even if our dataset ground truth annotations were performed according to the COCO keypoint detection task [20], its respective evaluation metric (AP, AP50, AP75) was not used since COCO AP is calculated using the Object Keypoint Similarity metric. As shown in equation (4), OKS uses per-keypoint constants that are calculated only from images in COCO validation set [10], which cannot be representative of another kind of datasets.

$$OKS = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right)\delta(\upsilon_i > 0)}{\sum_i \delta(\upsilon_i > 0)} \tag{4}$$

where d_i is the Euclidean distance between the detected keypoint and the corresponding ground truth, υ_i is the visibility flag of the ground truth, s is the object scale, and k is a per-keypoint constant that controls falloff.

The Percentage of Correct Key-points (PCK) is more suitable for infants' 2D pose estimation evaluation. It defines a correctly detected joint if the distance between the predicted joint location and the ground truth is smaller than a certain threshold defined by a fraction of torso size. For this work, the torso height (the distance between the right hip and right shoulder) was used instead of torso diagonal size (the distance between the right hip and left shoulder) since torso diagonal size can vary depending on the infant's pose as can be seen in Fig. 7. The notations PCK@0.2, PCK@0.1, and PCK@0.05 were used to refer to the fractions of torso height used as thresholds 0.2, 0.1, and 0.05 respectively (5). For 3D pose evaluation, 3DPCK was used. It defines the keypoint as correct if it falls within a given distance from the ground truth: 1cm, 2.5 cm, 5cm, and 10cm. These thresholds can show how accurate the keypoints estimations are at different levels. For clinical usage and considering that the average infant height is 50 centimeters, a 2.5 cm error will represent 5% of the body size which is acceptable. But errors that exceed this value (5cm, 10cm) are considered as too important. On the other hand, errors that do not exceed 1cm can be tolerated for this use case.

$$PCK@X = \frac{\sum_{i=1}^{N} \delta(d_i^2 \leq X \times T)}{N} \tag{5}$$

where d_i is the Euclidean distance between the detected keypoint and the corresponding ground truth, X is the threshold fraction, T is the torso height and N the total number of keypoints.

4 Experiments

The goal of this section is twofold: First, to compare the different selected deep learning architectures and their retrained versions on our dataset. Second, to quantify the precision and accuracy of the keypoint detection in the perspective of clinical use.

The OpenPose network which was originally trained on adult human pose estimation, is retrained by Chambers et al. [7] on an infants' dataset of 9039 images collected mainly from Youtube. This retrained network is widely used to automatically assess infants' movements. So we compare and test this model without further training on our images and check whether it can generalize well on data from a clinical environment only. The DarkPose + FiDIP network [16] was a suggested solution to solve the lack of clinical data. Based on a hybrid dataset of synthetic and real images collected from the internet (SyRip), a domain classifier was trained to promote a feature extractor to retain the ability to extract keypoints' information but also to ignore the differences between the real and the synthetic input images. This network is also tested and compared without extra training as an existing solution in addition to OpenPose retrained version.

The experiments in [38] have shown that HRNet networks with higher input resolution (384 × 288) improve the AP compared to networks with lower input resolution (256 × 192). For that reason, two HRNet trained networks (W32 & W48) are chosen with an input resolution of 384 × 288, in addition to two HigherHRNet trained networks (W32 & W48) with an input resolution of 640 × 640, and finally two DarkPose on top of HRNet trained networks (W32 & W48) with input resolution of 384 × 288. These networks are tested with and without training on our infants' clinical dataset. For training, we used only 15 epochs since we are only adjusting the network weights. An Adam optimizer was used, and

the learning rate was set to 10^{-4} throughout the whole process. For data augmentation, the random scale was set to ([0.9, 1.1]), random rotation to ([−45°, 45°]), and without half-body data augmentation.

5 Results and discussion

5.1 2D pose estimation

Figure 6 shows the results obtained after training and testing the different architectures. The error is calculated in pixels as the Euclidean distance between ground truth and predicted keypoints and averaged for each image (36k images with 880x720 pixels resolution). The impact of retraining the networks on our infants' dataset is obvious, as it led to a noticeable decrease in the mean error across all architectures, as well as a reduction in the standard deviation. Specifically, for the HigherHRNet32 and HigherHRNet48 networks, the mean error diminished from 30.7 to 3.7 pixels and from 28.5 to 5.3 pixels respectively.

The two networks OpenPose and FiDIP, have been evaluated without any additional training. With a confidence score threshold of 0.1, it was observed that OpenPose exhibited the lowest mean error of 6 pixels (see Fig. 6). This result is remarkably low; however, it should be noted that the mean error calculation exclusively considered successfully predicted joints across the entire dataset and the corresponding mean detection rate for this network was approximately 5 keypoints per image, in contrast to the other networks that achieved a detection rate of 17 keypoints per image (100% of joints detected). This discrepancy becomes evident when assessing the network's performance using PCK evaluation at various thresholds, as illustrated in Table 3. In fact, OpenPose performed poorly in comparison to networks that were not trained on any infants' images.

As referred in Table 3, the FiDIP network exhibited a comparable PCK (Percentage of Correct Keypoints) when compared to the non-retrained DarkPose48 network (95.35% versus

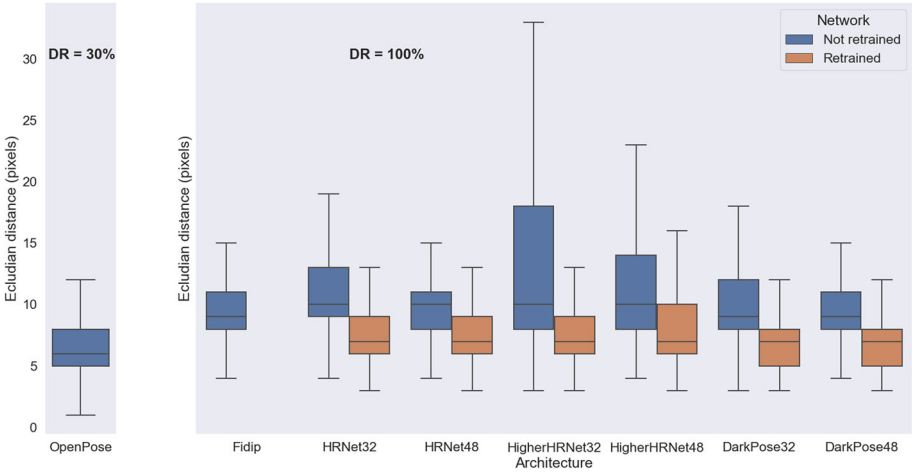

Fig. 6 2D pose estimation error for different networks in pixels. The boxplot shows the median error and the whiskers mark 1.5 times the interquartile range. DR refers to detection rate

Table 3 2D pose estimation results comparison of original *versus* retrained networks on our test dataset

Network	Original			Retrained		
	PCK@0.05	PCK@0.1	PCK@0.2	PCK@0.05	PCK@0.1	PCK@0.2
Openpose	50.08	60.31	64.07	–	–	–
FiDIP	55.93	84.29	95.35	–	–	–
HRNet32	52.31	79.95	92.41	66.35	90.92	97.98
HRNet48	55.86	83.49	95.01	65.84	90.82	98.25
Higher32	53.80	75.58	85.24	65.90	89.47	97.44
Higher48	54.89	77.07	87.78	65.16	88.27	96.41
DarkPose32	54.44	82.09	93.31	68.07	91.37	98.30
DarkPose48	56.58	84.39	95.32	66.93	90.97	98.28

95.32% for PCK@0.2) . Despite being retrained using the SyRIP dataset, which comprises 700 real and 1000 synthetic images, the FiDIP network did not demonstrate any advantage over DarkPose48. This lack of improvement can be attributed to the characteristics of the images within this dataset (refer to Fig. 4 and Table 2), which were sourced from the internet and encompass infant poses that do not adequately represent the typical video acquisition protocol for assessing infants. Consequently, retraining on synthetic images does not inherently enhance the efficiency of networks when processing real images captured in a clinical setting. This underscores the need for a real and significant dataset in this particular research domain, which cannot be overlooked or substituted.

The HigherHRNet networks demonstrated a satisfactory level of performance in terms of PCK, but not as good as HRNet networks. This disparity can be attributed to the fact that the scale of our subjects within the image is already adequate. Consequently, upsampling the predicted heatmaps and averaging them across all scales may potentially impact the accuracy of the predictions. This observation aligns with the findings of Cheng et al. [8], who concluded that HigherHRNet performs exceptionally well when dealing with small scales. Furthermore, it is worth noting that the HigherHRNet32 network exhibited a notable standard deviation of error. Upon closer examination, it was identified that in certain test images, the network exhibited confusion between keypoints located on the left and right sides. However, this issue was successfully addressed through network retraining, as it was no longer observed thereafter.

Both architectures of DarkPose (W32 & W48) achieved the best PCK (91.37% & 90.97% at 0.1 threshold and 98.30% & 98.28% at 0.2 threshold respectively). This proves that instead of using the standard coordinate decoding method, the heatmap decoding process described in Section 3.2.3 can improve infants' pose accuracy, which is in accordance with the results in [41]. Another observation is that after training the W32 and W48 versions, they have approximately the same performance, and this is because our subject's bounding boxes are large enough, so a lighter version of these networks can be used to reduce time complexity and resources.

5.2 3D pose estimation

The 3D pose estimation results were in accordance with the results obtained in 2D analysis (see Table 4). Openpose had the least 3DPCK at all thresholds with the same detection rate

Table 4 3D pose estimation results comparison of the non-retrained networks on our test dataset

Original	Openpose	FiDIP	HRNet32	HRNet48	Higher32	Higher48	DarkPose32	DarkPose48
3DPCK@1cm	17.46	32.71	26.46	28.72	31.87	32.63	31.81	32.97
3DPCK@2.5cm	40.75	76.26	67.47	72.31	65.94	68.06	73.64	76.46
3DPCK@5cm	53.20	93.88	88.63	92.56	81.03	84.08	91.20	93.86
3DPCK@10cm	57.12	98.38	95.80	98.00	87.94	90.12	96.96	98.20
Mean error (cm)	2.44	2.31	3.77	2.62	8.96	8.52	2.97	2.37
STD (cm)	5.60	8.06	24.39	6.91	70.62	72.18	16.81	8.81
Retrained	–	–	HRNet32[1]	HRNet48[1]	Higher32[1]	Higher48[1]	Dark32[1]	Dark48[1]
3DPCK@1cm	–	–	35.43	35.33	39.33	37.56	40.49	40.24
3DPCK@2.5cm	–	–	79.47	79.99	80.48	78.35	83.11	82.92
3DPCK@5cm	–	–	96.66	96.74	95.75	93.93	97.29	97.26
3DPCK@10cm	–	–	99.33	99.43	98.69	97.79	99.45	99.45
Mean error (cm)	–	–	1.91	1.90	1.98	2.36	1.72	1.73
STD (cm)	–	–	2.50	4.48	7.34	10.39	2.31	2.34

[1] Retrained on our infants' images dataset

Table 5 Per joints 3D pose estimation results of the retrained DarkPose32 network

	Nose	Shoulders	Elbows	Wrists	Hips	Knees	Ankles
3DPCK@1cm	65.38	27.57	42.19	48.70	15.12	44.34	35.36
3DPCK@2.5cm	95.41	81.61	86.78	87.28	66.93	88.68	82.99
3DPCK@5cm	99.74	98.25	98.55	98.07	95.08	98.76	98.00
3DPCK@10cm	99.99	99.98	99.88	99.58	99.94	99.93	99.57
Mean error (cm)	0.94	1.72	1.45	1.42	2.24	1.39	1.71
STD (cm)	0.96	1.11	1.14	1.59	1.42	1.07	3.73

of 30%, and FiDIP had comparable results with not retrained DarkPose48 (93.88% versus 93.86% at 3DPCK@5cm) despite being retrained using the SyRIP dataset, which validates that training on artificial images alone does not automatically improve the effectiveness of networks.

All the retrained networks gained 5% minimum of 3DPCK@2.5cm, and their mean error was reduced. The retrained version of DarkPose32 achieved the best 3DPCK at all thresholds, with a minimum mean error of 1.72 cm, which is a very promising result regarding the complex poses in our testing dataset. The results for DarkPose48 network after training were very similar to the 32 version with only a 3mm difference in mean error.

Moreover, when evaluating the results of this latter network for each group of joints separately, the mean error can be even lower, particularly for keypoints that are easy to define, such as the nose for which the retrained DarkPose32 network achieved 99.74% in 3DPCK@2.5cm and a mean error of 0.94 cm (see Table 5). The same table shows how higher the error is for keypoints that are difficult to localize precisely, such as the hips, for which the mean error is more than 2cm. We can also notice that better 2D pose detections lead to better 3D pose estimations, as is the case for DarkPose Networks, and the opposite is valid also since HigherHRNet had a considerable standard deviation and mean error.

Figure 7 shows the results of some networks on different complex infants' poses. It can be observed that all three networks before training do not provide accurate 3D pose estimations, which become better after training even when joints are half or completely occluded, as is the case for the second image. A standard depth camera will not be able to estimate the depth of these joints since it represents visible body surface depth information only, which shows the benefits and advantages of using our framework to analyze infants' movements, as can be seen in the video Online Resource 1.

In summary, the main conclusions that can be drawn from these results are: (1) it is particularly important to use a dedicated annotated dataset to train pose detection models for infants in a real medical environment; (2) compared to existing retrained models Openpose and FiDIP, models built on the HRNet network provide better precision, particularly DarkPose32 which is the best performing one in our case; (3) the latter network provides a 3D localization error of 1.7cm with a 3DPCK at 2.5cm of 83% and a 3DPCK at 5cm of 97%.

6 Conclusion

The 3D automatic GMA method is a challenging field of study. An accurate estimation of infant poses is needed to automatically analyze the complexity, variability, and fluidity of infants' movements. This study is the first stereoscopic 3D infant's pose estimation framework

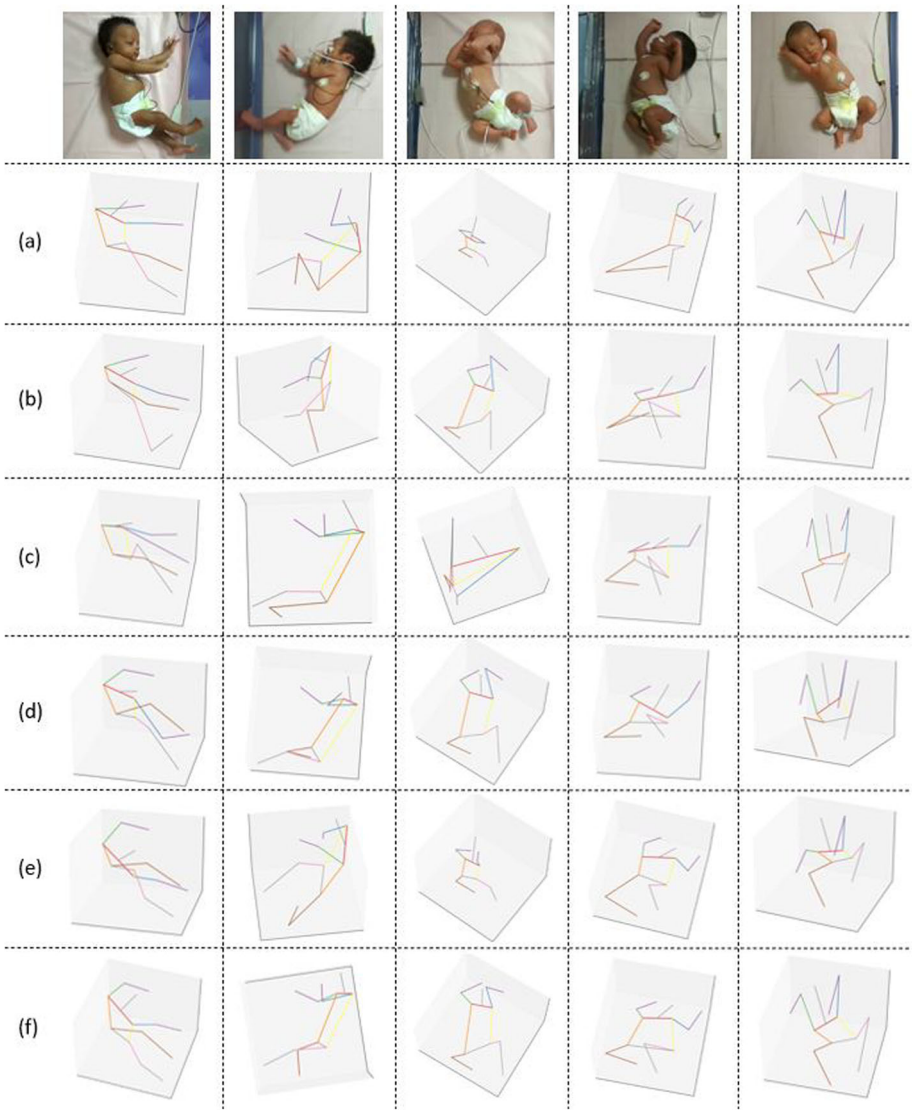

Fig. 7 3D pose estimation results of different networks. (a) HRNet32, (b) retrained HRNet32, (c) HigherHRNet32, (d) retrained HigherHRNet32, (e) DarkPose32, (f) retrained DarkPose32

dedicated to preterm children hospitalized in the Neonatal Unit. Compared to other existing automatic infant movement analysis tools, AGMA stereoscopic framework has successfully shown the possibility of estimating accurate 3D infant poses without the use of any markers or infrared cameras. Three state-of-the-art 2D human pose estimation networks (HRNet, HigherHRNet, and DarkPose) were retrained on a dataset of 88,500 preterm infant images collected in a real medical environment and manually annotated. The networks were tested with 18000 stereoscopic images and compared to the latest works on 2D infant pose estimation retrained on real or synthetic images. The minimum mean 3D error on joint position achieved

was 1.7 cm with a 3DPCK at 5cm of 97%. This study demonstrated that retraining on synthetic images did not consequently make networks efficient on real images captured in a clinical environment and that an adapted real dataset was more advantageous. In addition, it showed that an appropriate heatmap decoding process could improve infants' pose accuracy instead of using the standard coordinate decoding method. Another advantage of the presented framework is that any other advanced deep neural network can be used in the future for 2D pose estimation and then 3D pose reconstruction. Beyond GMA, AGMA 3D framework could also pave the way for the development of new tools based on the analysis of preterm infant movements, such as the automation of clinical seizure detection [28] and sleep quantification [5].

Author Contributions All authors substantially contributed to the conception and design, the acquisition of data, or the analysis and interpretation of data. AS drafted the manuscript. All authors revised the manuscript critically for important intellectual content and approved the final version to be published.

Funding AS was funded by a *Contrat Doctoral de l'École Doctorale 488 SIS*.

Availability of data and materials The retrained networks are available on https://drive.google.com/drive/folders/1GWKOLzGCwmontAUCrn8CoS3grQCUUqhE?usp=sharing

Code Availability Not applicable.

Declarations

Ethics approval The study was conducted in accordance with international ethical standards and the Declaration of Helsinki. The AGMA study was approved by the Comité de Protection des Personnes - Sud-Est II Ethical Committee in February 2021.

Consent to participate Written parental consent was obtained from each participant.

Consent for publication All the infants' images in this study were used after a signed parental consent.

Conflicts of interest The authors declare no competing interests.

References

1. Adde L, Helbostad JL, Jensenius AR, Taraldsen G, Støen R (2009) Using computer-based video analysis in the study of fidgety movements. Early Human Develop 85(9):541–547
2. Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2d human pose estimation: New benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 3686–3693
3. Baccinelli W, Bulgheroni M, Simonetti V, Fulceri F, Caruso A, Gila L, Luisa Scattoni M (2020) Movidea: A software package for automatic video analysis of movements in infants at risk for neurodevelopmental disorders. Brain Sciences, 10

4. Berthouze L, Mayston M (2011) Design and validation of surface-marker clusters for the quantification of joint rotations in general movements in early infancy. J Biomech 44(6):1212–1215
5. Cabon S, Weber R, Cailleau L, Carrault G, Pladys P, Porée F (2021) Automated quiet sleep detection for premature newborns based on video and ecg analysis. In: 2021 Computing in Cardiology (CinC), vol 48, pp 1–4
6. Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y (2021) Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 43(01):172–186
7. Chambers C, Seethapathi N, Saluja R, Loeb H, Pierce S, Bogen D, Prosser L, Johnson M, Kording K (2020) Computer vision to automatically assess infant neuromotor risk. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28:2431–2442
8. Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L (2020) Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5385–5394
9. Cioni G, Ferrari F, Bos AF, Prechtl HFR, Einspieler C (2008) Prechtls Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants. Mac Keith Press
10. Coco keypoint evaluation. https://cocodataset.org/#keypoints-eval
11. Doroniewicz I, Ledwoń D, Affanasowicz A, Kieszczyńska K, Latos D, Matyja M, Mitas A, Myśliwiec A (2020) Writhing movement detection in newborns on the second and third day of life using pose-based feature machine learning classification. Sensors, 20(21), 5986
12. Fan M, Gravem D, Cooper DM, Patterson DJ (2012) Augmenting gesture recognition with erlang-cox models to identify neurological disorders in premature babies. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12, page 411-420, New York, NY, USA. Association for Computing Machinery
13. Gravem D, Singh M, Chen C, Rich J, Vaughan J, Goldberg K, Waffarn F, Chou P, Cooper D, Reinkensmeyer D, Patterson D (2012) Assessment of Infant Movement With a Compact Wireless Accelerometer System. J Med Devices 6(2):021013
14. Groos D, Ramampiaro H, Ihlen E (2021) Efficientpose: Scalable single-person pose estimation. Appl Intell 51:2518–2533
15. Hesse N, Bodensteiner C, Arens M, Hofmann UG, Weinberger R, Schroeder AS (2019) Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In: Leal-Taixé L, Roth S, (eds), Computer Vision – ECCV 2018 Workshops, pp 32–49, Cham. Springer International Publishing
16. Huang X, Fu N, Liu S, Ostadabbas S (2021) Invariant representation learning for infant pose estimation with small data. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp 1–8, Los Alamitos, CA, USA. IEEE Computer Society
17. Karch D, Kim K-S, Wochner K, Pietz J, Dickhaus H, Philippi H (2008) Quantification of the segmental kinematics of spontaneous infant movements. J Biomech 41(13):2860–7
18. Karch D, Wochner K, Kim K, Philippi H, Hadders-Algra M, Pietz J, Dickhaus H (2010) Quantitative score for the evaluation of kinematic recordings in neuropediatric diagnostics detection of complex patterns in spontaneous limb movements. Methods Inform Med 49:526–530
19. Li M, Wei F, Li Y, Zhang S, Xu G (2021) Three-dimensional pose estimation of infants lying supine using data from a kinect sensor with low training cost. IEEE Sens J 21(5):6904–6913
20. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds), Computer Vision – ECCV 2014, pp 740–755, Cham. Springer International Publishing
21. McCay KD, Ho ESL, Shum HPH, Fehringer G, Marcroft C, Embleton ND (2020) Abnormal infant movements classification with deep learning on pose-based features. IEEE Access 8:51582–51592
22. McCay KD, Hu P, Shum HPH, Woo WL, Marcroft C, Embleton ND, Munteanu A, Ho ESL (2022) A pose-based feature fusion and classification framework for the early prediction of cerebral palsy in infants. IEEE Trans Neural Syst Rehab Eng 30:8–19
23. Meinecke L, Breitbach-Faller N, Bartz C, Damen R, Rau G, Disselhorst-Klug C (2006) Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy. Human Movement Sci 25(2):125–144
24. Moccia S, Migliorelli L, Pietrini R, Frontoni E (2019) Preterm infants' limb-pose estimation from depth images using convolutional neural networks. In: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp 1–7
25. Mohan S, Patterson DJ (2010) Involuntary gesture recognition for predicting cerebral palsy in high-risk infants. In: International Symposium on Wearable Computers (ISWC) 2010:1–8
26. Olsen JE, Cheong JLY, Eeles AL, FitzGerald TL, Cameron KL, Albesher RA, Anderson PJ, Doyle LW, Spittle AJ (2020) Early general movements are associated with developmental outcomes at 4.5-5 years. Early Human Develop 148:105115

27. Orlandi S, Raghuram K, Smith CR, 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), pp 3598–3601
28. Pavel A, Rennie J, Vries L, Blennow M, Foran A, Shah D, Pressler R, Kapellou O, Dempsey E, Mathieson S, Pavlidis E, Huffelen A, Livingstone V, Toet M, Weeke L, Finder M, Mitra S, Murray D, Marnane W, Boylan G (2020) A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. The Lancet Child & Adolescent Health 4:08
29. Pierrat V, Marchand-Martin L, Marret S, Arnaud C, Benhammou V, Cambonie G, Debillon T, Dufourg M-N, Gire C, Goffinet F, Kaminski M, Lapillonne A, Morgan AS, Rozé J-C, Twilhaar S, Charles M-A, Ancel P-Y (2021) Neurodevelopmental outcomes at age 5 among children born preterm: Epipage-2 cohort study. BMJ, 373
30. Raghuram K, Orlandi S, Shah V, Chau T, Luther M, Banihani R, Church P (2019) Automated movement analysis to predict motor impairment in preterm infants: a retrospective study. J Perinatology 39:1–8
31. Rahmati H, Aamo OM, Stavdahl Ã, Dragon R, Adde L (2014) Video-based early cerebral palsy prediction using motion segmentation. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 3779–3783
32. Reich S, Zhang D, Kulvicius T, Bölte S, Nielsen-Saines K, Pokorny F, Peharz R, Poustka L, Wörgötter F, Einspieler C, Marschik P (2021) Novel ai driven approach to classify infant motor functions. Scientific Reports, 11
33. Sakkos D, McCay K, Marcroft C, Embleton N, Chattopadhyay S, Ho E (2021) Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy. IEEE Access, PP:1–1
34. Sciortino G, Farinella GM, Battiato S, Leo M, Distante C (2017) On the estimation of children's poses. In: Battiato S, Gallo G, Schettini R, Stanco F (eds), Image Analysis and Processing - ICIAP 2017, pp 410–421, Cham. Springer International Publishing
35. Shivakumar SS, Loeb H, Bogen DK, Shofer F, Bryant P, Prosser L, Johnson MJ (2017) Stereo 3d tracking of infants in natural play conditions. In: 2017 International Conference on Rehabilitation Robotics (ICORR), pp 841–846
36. Spittle A, Orton J, Anderson P, Boyd R, Doyle L (2015) Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. Cochrane Database of Systematic Reviews
37. Stahl A, Schellewald C, Stavdahl Ã, Aamo O, Adde L, Kirkerød, H (2012) An optical flow-based method to predict infantile cerebral palsy. IEEE Transactions on Neural Systems and Rehabilitation Engineering: a Publication of the IEEE Engineering in Medicine and Biology Society, 20:605–614
38. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5686–5696
39. Wu Q, Xu G, Fan W, Chen L, Sicong Z (2021) Rgb-d videos-based early prediction of infant cerebral palsy via general movements complexity. IEEE Access PP:1–1
40. Wu Q, Xu G, Zhang S, Li Y, Wei F (2020) Human 3d pose estimation in a lying position by rgb-d images for medical diagnosis and rehabilitation. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp 5802–5805
41. Zhang F, Zhu X, Dai H, Ye M, Zhu C (2020) Distribution-aware coordinate representation for human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7091–7100