2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Asymmetric Three-dimensional Convolutions For Preterm Infants' Pose Estimation

Lucia Migliorelli[1,*], Daniele Berardini[1], Francesca Rossini[1], Emanuele Frontoni[1], Virgilio Carnielli[2] and Sara Moccia[3]

Abstract— **Computer-assisted tools for preterm infants' movement monitoring in neonatal intensive care unit (NICU) could support clinicians in highlighting preterm-birth complications. With such a view, in this work we propose a deep-learning framework for preterm infants' pose estimation from depth videos acquired in the actual clinical practice. The pipeline consists of two consecutive convolutional neural networks (CNNs). The first CNN (inherited from our previous work) acts to roughly predict joints and joint-connections position, while the second CNN (Asy-regression CNN) refines such predictions to trace the limb pose. Asy-regression relies on asymmetric convolutions to temporally optimize both the training and predictions phase. Compared to its counterpart without asymmetric convolutions, Asy-regression experiences a reduction in training and prediction time of 66% , while keeping the root mean square error, computed against manual pose annotation, merely unchanged. Research mostly works to develop highly accurate models, few efforts have been invested to make the training and deployment of such models time-effective. With a view to make these monitoring technologies sustainable, here we focused on the second aspect and addressed the problem of designing a framework as trade-off between reliability and efficiency.**

I. Introduction

Full-term pregnancy is defined by the World Health Organization (WHO) as a birth between 37 and 42 weeks of gestation. Any birth before the 37 gestational weeks is known as preterm birth. Preterm birth is a major global health issue, being responsible for the majority of motor, visual and learning disabilities in children and young adults [1].

Preterm infants are admitted immediately after birth into Neonatal Intensive Care Unit (NICUs) as they are not fully developed, weak and have fluctuating vital signs [1]. In the NICU, clinicians pay particular attention in monitoring preterm infants' general movements (GMs). Infants whose GMs are absent or abnormal are at higher risk of developing cerebral palsy (CP) [2].

Infants' movement assessment is today performed with the qualitative and sporadic observations by trained clinicians of the infants' limbs directly at the crib. An automatic, objective and continuous measurement of spontaneous infants' movement could provide a better understanding of infants' health status and may reveal the presence of relevant pathology in advance [3].

Fig. 1. Depth-image acquisition setup: the RGB-D camera is positioned at approximately 40 cm over the infant's crib.

A. Related work and main contribution

A number of approaches to quantitatively measure preterm infants' spontaneous motility has been proposed in the literature. Wearable sensors, in [5], [6], are exploited for limb-movement detection. However, sensors may add an additional burden to infants', causing pain and discomfort.

A valuable alternative to wearable sensors is to use non-obstructive monitoring systems (e.g., RGB or RGB-D cameras) [7]. In [8], [9] algorithms for whole-body movement detection were implemented. However, monitoring each limb individually is crucial to highlight GMs impairments.

With the perspective of quantitatively monitoring preterm infants' single-limb movement, in our previous work [4], we implemented a deep learning (DL) framework. The framework processes depth video-clips acquired directly in NICU (Fig. 1) with two consecutive CNNs, followed by a joint-linking step. The first CNN (detection CNN) roughly detects joint and joint-connection, while the second (regression

This work was supported by the European Union through the grant SINC - System Improvement for Neonatal Care under the EU POR FESR 14-20 funding program.

[1]L. Migliorelli, F. Rossini, D. Berardini and E. Frontoni are with the Department of Information Engineering, Università Politecnica delle Marche, Italy

[2]V. Carnielli is with the Department of Neonatology, University Hospital Ancona, Università Politecnica delle Marche, Italy

[3]S. Moccia is with The BioRobotics Institute, Scuola Superiore Sant'Anna and with the Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Italy

*Correspondence to L. Migliorelli: l.migliorelli@pm.univpm.it

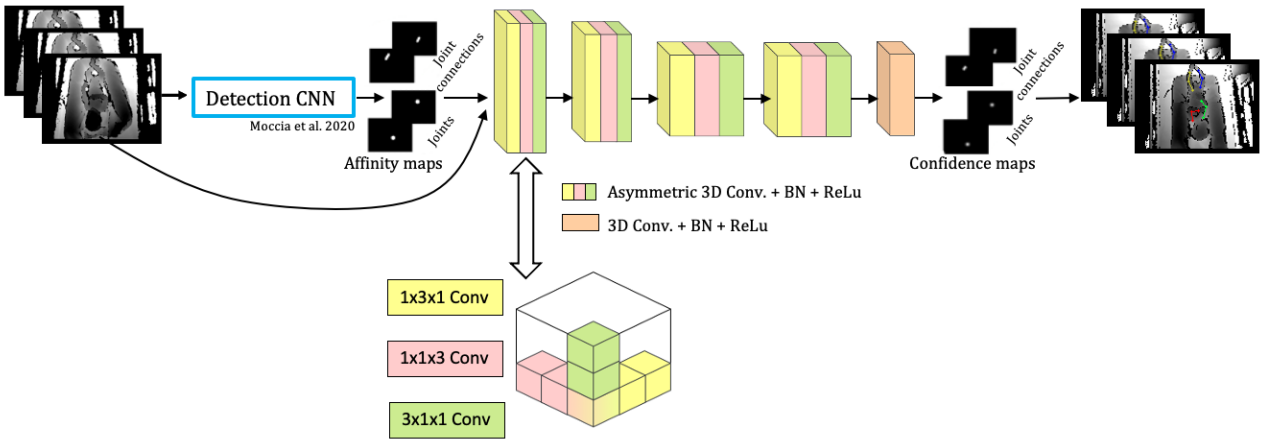

Fig. 2. Workflow of the proposed framework to preterm infants' pose estimation with spatio-temporal features extracted from depth video clips. The input consists of a temporal clip of 3 consecutive depth frames, which are processed by two convolutional neural networks (CNNs). The first CNN (which is inherited from our previous work [4]) roughly detects joint and joint-connection while the regression CNN (Asy-regression), which implements asymmetric 3D convolutions, acts to refine joint and joint-connection detection (confidence maps). Batch normalization (BN) and activation with rectified linear unit activation (ReLu) are implemented after each convolution (Conv).The detail of the asymmetric convolution expressed as 3 cascaded convolutions (1x3x1, 1x1x3, 3x1x1) is shown at the bottom.

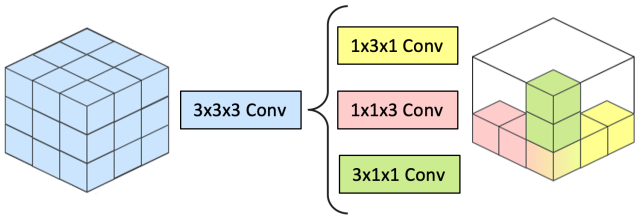

Fig. 3. Approximation of a three dimensional (3D) convolutional layer with kernel size of 3x3x3 (blue) by three cascaded asymmetric 3D convolutional layers with kernel sizes of 1×3×1 (yellow), 1×1×3 (pink) and 3×1×1 (green).

CNN) precisely regresses joint and joint-connection position. The output of the regression CNN acts as guidance to trace limb pose. Such a framework, despite being robust in term of performance, naturally relies on a high number of parameters to process spatio-temporal features. This consequently increases both training and prediction time with a negative impact in terms of sustainability [10].

To simultaneously accomplish efficiency while attaining accuracy, in this work, we leverage asymmetric convolutions [11] in the regression CNN. The traditional three dimensional (3D) convolutions are here split into three cascaded asymmetric one-directional 3D convolutions with the same size of receptive field as the traditional 3D convolution. In this work, we focus on the asymmetric version of the regression CNN (Asy-Regression). Indeed, implementing three cascaded convolutions in the detection CNN would make the networks even deeper, posing issues related to overfitting or vanishing gradient [12].

II. Methods

The workflow of the proposed approach in shown in Fig. 2 while a graphical representation of the asymmetric convolutions is shown in Fig. 3.

TABLE I

Asy-Regression CNN architecture with the three cascaded asymmetric convolutions

Name	Kernel (Size / Stride)	Channels
Input	–	3x21
Layer 1	1x3x1 / 1x1x1	3x64
Layer 1	1x1x3 / 1x1x1	3x64
Layer 1	3x1x1 / 1x1x1	3x64
Layer 2	1x3x1 / 1x1x1	3x128
Layer 2	1x1x3 / 1x1x1	3x128
Layer 2	3x1x1 / 1x1x1	3x128
Layer 3	1x3x1 / 1x1x1	3x256
Layer 3	1x1x3 / 1x1x1	3x256
Layer 3	3x1x1 / 1x1x1	3x256
Layer 4	1x3x1 / 1x1x1	3x256
Layer 4	1x1x3 / 1x1x1	3x256
Layer 4	3x1x1 / 1x1x1	3x256
Layer 5	1x1x1 / 1x1x1	3x256
Output	1x1x1 / 1x1x1	3x20

A. *Infant's limb model and ground truth*

As in [4], to train the two consecutive CNNs (i.e., the detection CNN and the Asy-regression), we adopted temporal clips. Each clip consisted of 3 consecutive depth frames. For each depth frame we constructed the associated 20 ground-truth masks (i.e., 12 for the limb-joint and 8 for the joint-connection). This approach with individual ground truth masks is robust to eventual joint-occlusions caused by the presence of operators or limb-self-occlusions. Binary masks were used to train the detection CNN. For each joint, we considered all the pixels lying within a circle of radius r centered at the manually annotated joint-site. Similarly, the ground-truth for the joint-connection, was the rectangular region with thickness r and centrally aligned with respect to the line linking the two subsequent joints.

Gaussian-distributed masks were used to train the Asy-regression CNN. For each joint, we considered a region

consisting of all pixels laying in the circle with radius r centered at the manual annotation site. Such region was the Gaussian distributed version of the binary mask with standard deviation equal to $3*r$. Similarly, rectangular ground-truth masks were generated for the joint-connection. The Gaussian-distributed version of the previously defined binary mask were depicted along the connection direction with a standard deviation equal to $3*r$.

B. Preterm infants pose estimation framework

As shown in Fig. 2, the proposed framework for limb-pose estimation has 2 main stages: the detection CNN which is the same in [4] and the Asy-regression. The detection CNN is the one presented in [4]. It is inspired by the classic encoder-decoder architecture of U-Net with a two-branch architecture aimed at processing joints and joint connections separately. The detection network is fed with a clip of 3 depth frames. For each depth frame in the clip, 20 binary ground-truth affinity maps are generated.

The architecture of the Asy-regression CNN presented in this work is reported in Table I. The CNN is fed by stacking the depth temporal clip and the corresponding affinity maps obtained from the detection CNN. In the first 3 layers, the number of activations is doubled, ranging from 64 to 256. The number of activations is then kept constant for the last two layers. In this sub-network, the 3D convolutional layer is approximated using asymmetric 3D convolutions, i.e., three cascaded asymmetric 3D convolutional layers with kernel sizes of 1×3×1, 1×1×3 and 3×1×1. As shown in Fig. 3, the three cascaded asymmetric 3D convolutional layers have same size of receptive field, as the traditional 3D convolutional layer, while decreasing the number of parameters and computational cost significantly [12].

In both the CNNs, batch normalization and activation with the rectified linear unit (ReLu) are performed after each convolutional layer. As in [4], the last step of the framework consists in linking subsequent joints to trace the skeleton of the limbs. This is a multi-stage approach: first, joint candidates are identified from the output joint-confidence maps using non-maximum suppression, then the candidates are linked exploiting the joint-connection confidence maps via a bipartite graph matching.

III. Experimental protocol

A. Dataset

For this work, we extended the Babypose dataset [13], which originally accounted for 16 depth videos, to have a total of 22 depth videos from 22, spontaneously breathing, preterm infants. The videos were acquired in the NICU of the G. Salesi Hospital in Ancona, Italy. The Astra Mini S Orbbec ®was used to record the videos. The camera has a frame rate of 30 frames per second with image size of 640x480 pixels. Each video is 180 s-long. Joint-annotation was performed using a custom-built annotation tool[1]. As in [4], for each of the 22 videos, 1 frame every 5 was extracted and 1000 frames were annotated per infant. Three subsequent frames were coupled to form a video clip. In the training set, subsequent depth clips were shifted by one frame (accounting for 751 depth clips per infants) while no shifting was performed in the testing set resulting in 83 depth clips.

[1]https://github.com/roccopietrini/pyPointAnnotator

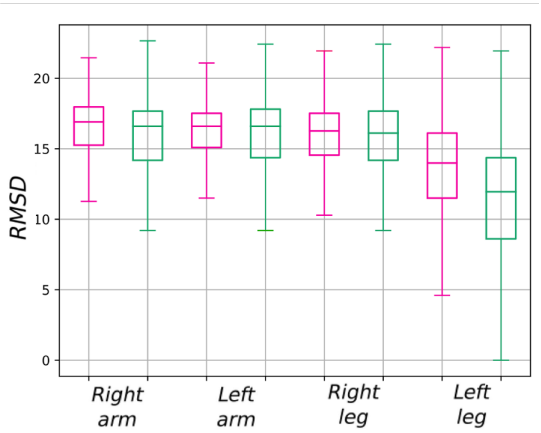

Fig. 4. Limb-pose estimation performance in terms of median root mean square distance (*RMSD*) computed with respect to the ground-truth pose. The *RMSD* is reported for each limb, separately. Results are reported for the Regression in [4] (green) and Asy-regression (pink)-based framework.

B. Training settings

To train the CNNs, each image in the depth clip was resized to 128x96 pixels. Mean was removed by each image in the clip. For each image we created the ground-truth masks with r equal to 6 pixels. For the detection CNN the per-pixel binary cross entropy was used as loss function and Adam as optimizer. The Asy-regression network was trained with the stochastic gradient descent (SGD) as optimizer (Momentum=0.98) using the mean squared error as loss function. Both the losses were adapted for multiple maps training. An initial learning rate of 0.01 was used with a learning decay of 10% every 10 epochs. The batch size was equal to 16 and the number of epochs equal to 150 for the detection network and to 100 for the Asy-regression one. We selected the best model among the epochs as the one that maximized the detection accuracy and minimized the mean absolute error on the validation set, for the detection and the Asy-regression CNN, respectively.

The training was performed on a Nvidia GeForce RTX 2080 11 GB.

C. Performance metrics and comparison with the literature

To evaluate the performance of the framework in estimating infants' limb pose, we computed the root mean square distance (*RMSD*) [pixels] for the 128x96 pixel images and for each infants' limb.

$$RMSD = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}} \quad (1)$$

TABLE II

Limb-pose estimation performance in terms of median RMSD computed with respect to ground-truth pose. IQR is reported in parentheses. The RMSD is reported separately for each limb. Results are reported for the traditional 3D and the asymmetric 3D (Asymm) frameworks. Testing time to predict a triplet of frames is shown too.

	Right arm	Left arm	Right leg	Left leg	Testing time
	RMSD				
Asy-regression	16.90 (2.71)	16.59 (2.43)	16.27 (2.97)	13.99 (4.60)	0.03
Regression in Moccia et al. [4]	16.59 (3.49)	16.59 (3.45)	16.10 (3.49)	11.95 (5.76)	0.05

where $\tilde{y}_i$ and y_i are the predicted and observed joint coordinates, respectively, and n is the number of observations. The Asy-regression was compared against its closer variant (i.e., the regression in [4]).

The $RMSD$ median values for pose estimation were calculated for both the Asy-regression and the regression in [4].

IV. Results

The achieved results in terms of $RMSD$ for both the two CNNs are shown in Table II. The $RMSD$ boxplot are shown in Fig. 4. These quantitative results showed that introducing the asymmetric convolutions kept the results in terms of error almost unchanged. Interquartile ranges (IQRs) was reported in Table II and were always lower than 4.60 pixel for the Asy-regression, while, for the regression in [4], were lower than 5.76. Thus, the introduction of the asymmetric convolution has increased the network generalization ability. Both the training and the prediction time was calculated to prove the effectiveness of the asymmetric convolutions in terms of temporal optimization. Asymmetric 3D convolution produced a reduction of the training time of 66%. The time to predict a depth-clip (Table II) was on average 0.03 s for the Asy-regression while for the regression in [4] was 0.05s.

V. Discussion and Conclusion

Monitoring preterm infants' movement in NICUs, through non-contact measures, is crucial for early assessing preterm-birth-related complications. In our previous work [4], we proposed a novel framework for non-intrusive monitoring of preterm infants' limbs. It provides an innovative approach for limb-pose estimation from spatio-temporal features extracted from depth video-clips. Although being robust, this approach was very parameter-intensive, resulting in too much time spent on training and testing phases. Searching for a trade-off between reliability and efficiency, in this work we implemented the Asy-regression network with asymmetric convolutions [12]. As showed in Sec. IV, such variation from the original version in [4], was able at lowering both the training and the prediction time of the 66% while keeping the $RMSD$ almost unchanged (mean $\Delta RMSD$ between the two architectures = 0.63 pixels).

Nowadays, researchers are all geared towards finding the most effective models, few effort is spent on making such models more efficient. Guided by these premises and the results achieved by this work, natural extension of the proposed approach would try to implement even lighter models [11]. This would enable the implementation of sustainable automated and intelligent monitoring systems in scenarios with fewer computational resources.

References

[1] M. Villarroel, S. Chaichulee, J. Jorge, S. Davis, G. Green, C. Arteta, A. Zisserman, K. McCormick, P. Watkinson, and L. Tarassenko, "Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–18, 2019.

[2] C. Morgan, C. Crowle, T.-A. Goyen, C. Hardman, M. Jackman, I. Novak, and N. Badawi, "Sensitivity and specificity of general movements assessment for diagnostic accuracy of detecting cerebral palsy early in an Australian context," *Journal of Paediatrics and Child Health*, vol. 52, no. 1, pp. 54–59, 2016.

[3] I. Zuzarte, A. H. Gee, D. Sternad, and D. Paydarfar, "Automated movement detection reveals features of maturation in preterm infants," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*. IEEE, 2020, pp. 600–603.

[4] S. Moccia, L. Migliorelli, V. Carnielli, and E. Frontoni, "Preterm infants' pose estimation with spatio-temporal features," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 8, pp. 2370–2380, 2020.

[5] C. B. Redd, L. A. Barber, R. N. Boyd, M. Varnfield, and M. K. Karunanithi, "Development of a wearable sensor network for quantification of infant general movements for the diagnosis of cerebral palsy," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2019, pp. 7134–7139.

[6] M. Airaksinen, O. Räsänen, E. Ilén, T. Häyrinen, A. Kivi, V. Marchi, A. Gallen, S. Blom, A. Varhe, N. Kaartinen *et al.*, "Automatic posture and movement tracking of infants with wearable movement sensors," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[7] K. Raghuram, S. Orlandi, P. Church, T. Chau, E. Uleryk, P. Pechlivanoglou, and V. Shah, "Automated movement recognition to predict motor impairment in high-risk infants: a systematic review of diagnostic test accuracy and meta-analysis," *Developmental Medicine & Child Neurology*, 2020.

[8] S. Orlandi, K. Raghuram, C. R. Smith, D. Mansueto, P. Church, V. Shah, M. Luther, and T. Chau, "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*. IEEE, 2018, pp. 3598–3601.

[9] Y. Sun, D. Kommers, W. Wang, R. Joshi, C. Shan, T. Tan, R. M. Aarts, C. van Pul, P. Andriessen, and P. H. de With, "Automatic and continuous discomfort detection for premature infants in a NICU using video-based motion analysis," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2019, pp. 5995–5999.

[10] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[11] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proceedings of the ACM Multimedia Asia*, 2019, pp. 1–6.

[12] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognition*, vol. 85, pp. 1–12, 2019.

[13] L. Migliorelli, S. Moccia, R. Pietrini, V. P. Carnielli, and E. Frontoni, "The babyPose dataset," *Data in Brief*, vol. 33, p. 106329, 2020.