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

Improving Preterm Infants' Joint Detection in Depth Images Via Dense Convolutional Neural Networks

Lucia Migliorelli[1,*], Emanuele Frontoni[1], Simone Appugliese[1],
Giuseppe Pio Cannata[1], Virgilio Carnielli[2] and Sara Moccia[3]

***Abstract*— Preterm infants' spontaneous motility is a valuable diagnostic and prognostic index of motor and cognitive impairments. Despite being recognized as crucial, preterm infant's movement assessment is mostly based on clinicians' visual inspection. The aim of this work is to present a 2D dense convolutional neural network (denseCNN) to detect preterm infant's joints in depth images acquired in neonatal intensive care units. The denseCNN allows to improve the performance of our previous model in the detection of joints and joint connections, reaching a median recall value equal to 0.839. With a view to monitor preterm infants in a scenario where computational resources are scarce, we tested the architecture on a mid-range laptop. The prediction occurs in real-time (0.014 s per image), opening up the possibility of integrating such monitoring system in a domestic environment.**

I. INTRODUCTION

The World Health Organization defines preterm infants as infants born before the thirty-seven weeks of gestation. Every year there are more than fifteen million worldwide preterm births and the number of cases is still growing. In almost all high-income countries, complications of preterm birth (such as brain and lungs complications) are the largest direct cause of neonatal deaths [1].

Infants' spontaneous motility has a valuable diagnostic and prognostic role [2]. However, movement evaluation is today mostly based on clinicians' visual inspections at the crib side in Neonatal Intensive Care Units (NICUs). This has the drawbacks of being qualitative, sporadic, and susceptible to intra- and inter-clinician variability.

To support clinicians in preterm infants' movement monitoring, computer-assisted solutions have been proposed in the years. In [3], [4], wearable sensors (i.e., accelerometers and gyroscopes) are used for limb-movement detection. The main limitation here is that wearable sensors may hinder infants spontaneous motility and further cause pain, discomfort and skin damage [5]. To overcome the drawbacks of wearable sensors, the authors in [6], [7] study infants' whole-body movement via RGB-D camera. However, monitoring each limb separately is essential to possibly assess the presence of brain lesions [8].

Fig. 1. Depth-image acquisition setup. The depth camera (white square) is positioned ∼40 cm over the infant's crib, hence leaving the healthcare operators free to move. A sample of depth image is shown on the upper right corner (light-blue square).

In our previous work [9], we proposed a workflow based on 2D convolutional neural networks (CNNs) to estimate preterm infant's limb pose from depth images acquired in the NICU of the "G. Salesi" Hospital in Ancona (Fig. 1). The workflow couples two consecutive CNNs: the first one (detection network) for roughly detecting joint and joint-connection and the second one (regression network) for limb-pose estimation. In [10], we showed that introducing the temporal dimension improves the performance over an analysis based on spatial features-only [9]. However, 3D convolution, which is needed for spatio-temporal processing, increases the number of CNN parameters, posing issues relevant to deployment in cost-effective devices.

As pointed out by a recent review in the field [5], there is the urgent need to work on developing new algorithms for the automatic identification of preterm infants' joints from video-data. Guided by this consideration, and aware of the limitations of our previous work [9], [10], in this work we propose a modified version of our 2D detection network presented in [9]. The new network (denseCNN) exploits dense block (DB)-based skip connections [11] to boost the detection network performance while keeping the computational requirement lower than our 3D detection CNN.

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, E. Frontoni, S. Appugliese and G.P. Cannata 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

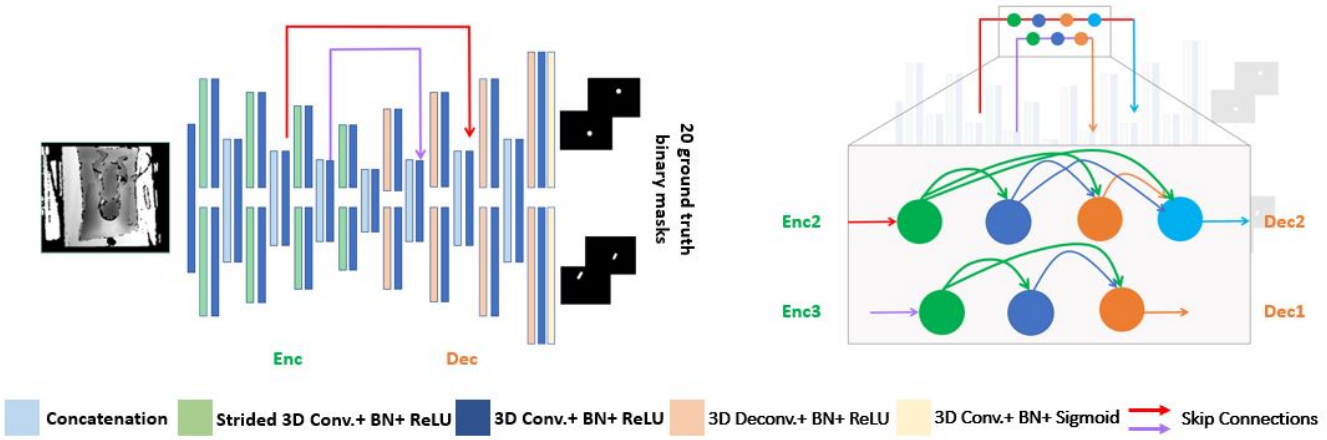

Fig. 2. *(Left) Graphic representation of the detection network from our previous work [9]. The network detects limb joints and joint connections. Batch normalization (BN) and activation with rectified linear unit activation (ReLu) are implemented after each convolutional layer. (Right) detail of the dense-block (DB)-based skip connections implemented on the denseCNN.*

II. METHODS

A. Ground-truth for joint and joint-connection

For training our denseCNN, we build 20 ground-truth binary masks, 12 for joints and 8 for joint connections. This choice is performed over having a single mask for all joints, due to: the presence of joint-occlusion (both from healthcare operators and caregiver and self-occlusion).

The mask for each joint consists of all pixels that lie inside the circle of a specific radius *(r)* that is centred in the centre of the joint. For the joint-connection, the ground truth consist in a rectangular region of a given thickness (equal to *r*) centrally aligned with the line that connects two consecutive joints.

B. denseCNN

Our denseCNN shares the same encoder-decoder architecture of the detection network in [9], with 8 convolutional blocks: 4 for the encoding path (downsampling) and 4 for the decoding path (upsampling) (Fig. 2). After the input layer, each block is divided in two branches for processing joint and joint-connection separately. The output of each single-branch is concatenated to enter in a single convolutional block which is newly separated in two branches. Two long-skip connections are added between the second encoder (Enc2) and second decoder (Dec2) block and the third encoder (Enc3) and first decoder block (Dec1).

The introduction of the skip connections between the encoding the decoding path is crucial to recover the spatial information lost during downsampling. However aggregating features from shallow (encoding) and deep (decoding) layers may pose issues relevant to semantic gap. Inspired by [11], to bridge the semantic gap induced by long-skip connections, we here implement DB-based skip connections (Fig. 2). In each DB, all layers are directly connected with each other: a layer got inputs from all preceding layers and distribute its own feature-maps to the next layers. Thus, the shallow feature maps from the encoding path are processed via the DB, which lowers their semantic order prior concatenating the deep features of the decoding path.

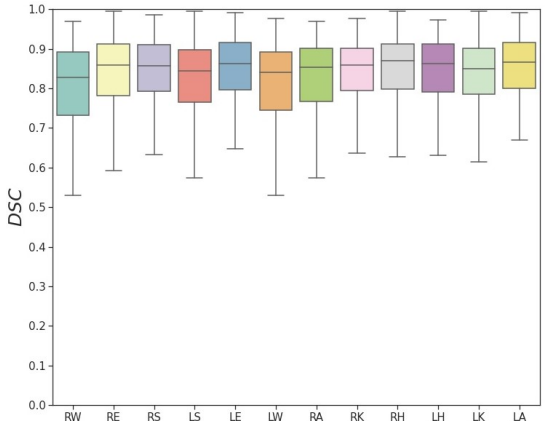

Fig. 3. *Boxplots of the Dice similarity coefficient (DSC) for joint detection achieved with the denseCNN.*

In this work, the DB between Enc2 and Dec2 has 4 layers with a growth rate of 64. The DB between Enc3 and Dec1 fuses features with less semantic distance thus it has only 3 layers with a growth rate of 128. Each dense layer implements 2 convolution operations. The former is a 3x3 convolution while the latter is a 1x1 convolution to the number of feature channels. In addition, a transition layer, which consists of 1x1 convolution is applied prior concatenating the decoder. Batch normalization and activation with the rectified linear unit (ReLu) are performed after each convolution.

III. EXPERIMENTAL PROTOCOL

A. Dataset

The dataset used in this work is an extended version of the BabyPose dataset [12]. Six additional depth videos were collected in the NICU of the "G. Salesi" Hospital (Ancona), after obtaining the written informed consent from the infants'

TABLE I

Joint-detection performance in terms of median recall. The metric is reported separately for each joint. LS and RS: left and right shoulder, LE and RE: left and right elbow, LW and RW: left and right wrist, LH and RH: left and right hip, LK and RK: left and right knee, LA and RA: left and right ankle.

	Right arm			Left arm			Right leg			Left leg		
	RW	RE	RS	LS	LE	LW	RA	RK	RH	LH	LK	LA
architecture in [9]	0.770	0.805	0.814	0.788	**0.850**	0.796	0.796	0.770	0.814	0.805	0.832	0.832
proposed denseCNN	**0.805**	**0.841**	**0.850**	**0.836**	0.841	**0.823**	**0.832**	**0.858**	**0.850**	**0.841**	0.832	**0.858**

TABLE II

Joint-connection detection performance in terms of median recall. The metric is reported separately for each joint-connection. LS and RS: left and right shoulder, LE and RE: left and right elbow, LW and RW: left and right wrist, LH and RH: left and right hip, LK and RK: left and right knee, LA and RA: left and right ankle.

	Right arm		Left arm		Right leg		Left leg	
	RW-RE	RE-RS	LS-LE	LE-LW	RA-RK	RK-RH	LH-LK	LK-LA
architecture in [9]	0.783	0.810	0.810	**0.854**	0.796	0.810	0.832	0.829
denseCNN	**0.812**	**0.852**	**0.838**	0.840	**0.833**	**0.844**	**0.835**	**0.851**

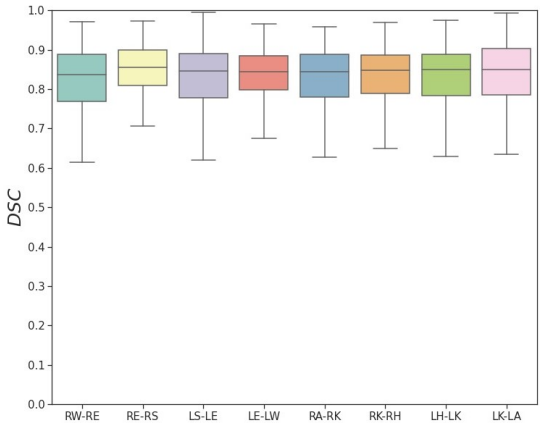

Fig. 4. *Boxplots of the Dice similarity coefficient (DSC) for joint-connection detection achieved with the denseCNN.*

legal guardians. The newly version of the BabyPose dataset accounts for 22 depth videos from 22 preterm infants.

Videos were recorded with the Astra Mini S-Orbbec®(frame rate = 30 frames per second, image size = 640x480 pixels). Joints were annotated under the supervision of an expert clinician. Considering the preterm infants' movement rate [13], annotation was performed every 5 frames until 1000 frames per infant were annotated. For each infant, 750 frames were used for training denseCNN and 250 to test it. This results in a training set of 16500 frames (of which, 5500 frames, were used for validating the architecture) and a test set of 5500 frames.

B. Training settings

The dataset images were resized to 128x96 pixels in order to smooth noise and reduce both training time and memory requirements. A joint radius r of 6 pixels was selected to build the ground-truth masks.

DenseCNN was trained for 100 epochs using the per-pixel binary cross entropy as loss function, and stochastic gradient descend as optimizer. An initial learning rate of 0.01 with a learning decay of 10% every 10 epochs was used, with 16 as batch size. All the analyses were performed using Keras on a Nvidia GeForce GTX 1050 Ti/PCIe/SSE2. Our model was tested on an Intel-Core i3-6006U @ 2.00GHz.

C. Performance metrics and comparison with the literature

Dice similarity coefficient (DSC) and recall (Rec) were computed to evaluate the performance of denseCNN:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

$$Rec = \frac{TP}{TP + FN} \tag{2}$$

where TP is the number of pixels correctly predicted as joint, and FP and FN are the number of misclassified pixels as belonging or not to the joint, respectively. By applying the same experimental protocol we compared the performance of denseCNN with the only work so far in literature to deal with depth-frame-based pose estimation (i.e., [9]).

IV. RESULTS

The median Rec for the 12 joints and the 8 connections, both for denseCNN and the network in [9], is reported in Table I and Table II, respectively. DenseCNN achieved a median *Rec* value of 0.839 and 0.838 for joints and joint-connections, respectively, overcoming the performance of the detection network in [9]. Figure 3 and Fig. 4 show the DSC boxplots of denseCNN, for the joint and joint-connection, respectively. The interquartile range (IQR) was always lower than 0.160 and 0.120 for joints and joint connections, respectively.

Figure 5 shows the qualitative results of the dense network. The network was able to correctly detect the joints also in challenging case, such as when a healthcare operator interacted with the infant or the infant crossed his/her legs. The prediction was performed with a cost-effective hardware (Intel-Core i3-6006U @ 2.00GHz). The dense network prediction time was of 0.014 s, allowing real-time processing.

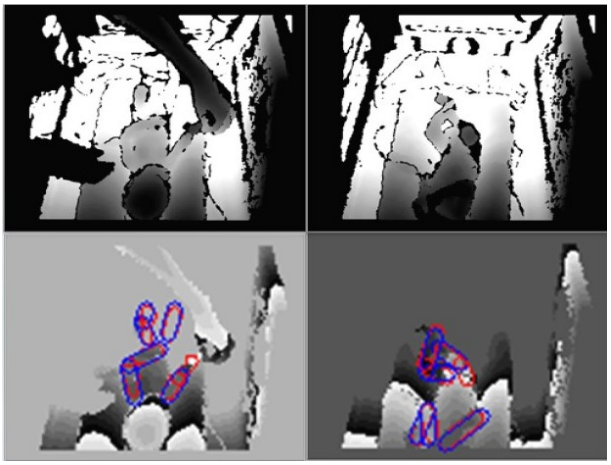

Fig. 5. *Sample detection results. First row: original depth image which shows challenges such as interaction with healthcare operator or presence of limbs self-occlusion. The Second row shows the predictions from the denseCNN with the preprocessed depth images superimposed with and the predictions (purple) and the ground truth (blue).*

V. DISCUSSION

Monitoring preterm infants' limb-movement in NICU is crucial for early recognizing preterm birth-related complications such as cerebral palsy. However, few approaches have been proposed to support clinicians in NICUs and there is still room for improvement [5]. In this paper, we presented an improvement of our previous work [9], implementing a dense architecture for limb-joint detection. The approach was validated on the second version of the BabyPose dataset, which consists of 22 depth video from 22 preterm infants.

As showed in Table I and Table II, denseCNN outperformed the network proposed in [9]. This support the hypothesis that introducing the DBs along the skip connections may effectively reduce the semantic gap enhancing the CNN ability to detect infants' joints. The DSC boxplots for both the joint (Fig. 3) and the joint-connection (Fig. 4) highlight the ability of network of performing comparably when detecting all joints and join-connections.

Qualitative results in Fig. 5 show the generalization ability of denseCNN even in challenging cases (e.g., presence of external occlusions or crossed legs). Further increasing the size of the BabyPose dataset with more complex cases (e.g., more videos in which the operators interact with the infants) would improve the robustness of this validation analysis and we are currently working in this direction.

As opposed to the approach in [10], denseCNN allowed real-time processing in cost-effective hardware, paving the way for the deployment in a domestic setting, where significant computing resources may not always be available.

VI. CONCLUSION

In this paper, we presented a dense architecture for preterm infants' joint and joint-connection detection in depth images acquired in NICUs. The network achieved encouraging results, suggesting that the introduction of the DBs along the skip connections may reduce the semantic gap between the encoder and decoder path [9].

With a view to continuously monitor preterm infants also in scenarios where computational resources are not always available, we deployed the model on a cost-effective hardware. The prediction time was compatible with real-time monitoring. These results open the possibility of translating such applications within the domestic environment.

REFERENCES

[1] A. Polito, S. Piga, P. E. Cogo, C. Corchia, V. Carnielli, M. Da Frè, D. Di Lallo, I. Favia, L. Gagliardi, F. Macagno *et al.*, "Increased morbidity and mortality in very preterm/VLBW infants with congenital heart disease," *Intensive Care Medicine*, vol. 39, no. 6, pp. 1104–1112, 2013.

[2] C. Einspieler, A. F. Bos, M. E. Libertus, and P. B. Marschik, "The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction," *Frontiers in Psychology*, vol. 7, p. 406, 2016.

[3] D. Gravem, M. Singh, C. Chen, J. Rich, J. Vaughan, K. Goldberg, F. Waffarn, P. Chou, D. Cooper, D. Reinkensmeyer *et al.*, "Assessment of infant movement with a compact wireless accelerometer system," *Journal of Medical Devices*, vol. 6, no. 2, 2012.

[4] I. A. Trujillo-Priego, C. J. Lane, D. L. Vanderbilt, W. Deng, G. E. Loeb, J. Shida, and B. A. Smith, "Development of a wearable sensor algorithm to detect the quantity and kinematic characteristics of infant arm movement bouts produced across a full day in the natural environment," *Technologies*, vol. 5, no. 3, p. 39, 2017.

[5] 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*.

[6] L. Adde, J. L. Helbostad, A. R. Jensenius, G. Taraldsen, and R. Støen, "Using computer-based video analysis in the study of fidgety movements," *Early Human Development*, vol. 85, no. 9, pp. 541–547, 2009.

[7] A. Cenci, D. Liciotti, E. Frontoni, A. Mancini, and P. Zingaretti, "Non-contact monitoring of preterm infants using RGB-D camera," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 57199. American Society of Mechanical Engineers, 2015, p. V009T07A003.

[8] D. Freymond, Y. Schutz, J. Decombaz, J.-L. Micheli, and E. Jéquier, "Energy balance, physical activity, and thermogenic effect of feeding in premature infants," *Pediatric Research*, vol. 20, no. 7, pp. 638–645, 1986.

[9] S. Moccia, L. Migliorelli, R. Pietrini, and E. Frontoni, "Preterm infants' limb-pose estimation from depth images using convolutional neural networks," in *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, 2019, pp. 1–7.

[10] 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.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

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

[13] B. Fallang, O. D. Saugstad, J. Grøgaard, and M. Hadders-Algra, "Kinematic quality of reaching movements in preterm infants," *Pediatric Research*, vol. 53, no. 5, p. 836, 2003.