3D Infant Pose Estimation Using Transfer Learning

Simon Ellershaw
Imperial College London
London, SW7 2BU, UK
simon.ellershaw19@imperial.ac.uk

Luca Schmidtke
Imperial College London
London, SW7 2BU, UK
l.schmidtke@imperial.ac.uk

Nidal Khatib
Imperial College London
London, SW7 2BU, UK
n.khatib@imperial.ac.uk

Jonathan Eden
Imperial College London
London, SW7 2BU, UK
j.eden@imperial.ac.uk

Jonathan Eden
Guy's and St Thomas' NHS Foundation Trust
London, SE1 9RS, UK
Anna.Jones1@gstt.nhs.uk

Sofia Dall'Orso
Imperial College London
London, SW7 2BU, UK
s.dallorso15@imperial.ac.uk

Silvia Muceli
Chalmers University of Technology
Gothenburg SE-412, Sweden
muceli@chalmers.se

Etienne Burdet
Imperial College London
London, SW7 2BU, UK
e.burdet@imperial.ac.uk

Niamh Nowlan
Imperial College London
London, SW7 2BU, UK
n.nowlan@imperial.ac.uk

Tomoki Arichi
King's College London
London, WC2R 2LS, UK
tomoki.arichi@kcl.ac.uk

Bernhard Kainz
Imperial College London
London, WC2R 2LS, UK
b.kainz@imperial.ac.uk

Abstract

This paper presents the first deep learning-based 3D infant pose estimation model. We transfer-learn models first trained in the adult domain. The model outperforms the current 2D and 3D state-of-the-art on the synthetic infant MINI-RGBD test dataset, achieving an average joint position error (AJPE) of 8.17 pixels and 28.47 mm respectively. Furthermore, unlike the current 3D state-of-the-art, the model presented here does not require a depth channel as input. This is an important step in the development of an automated general movement assessment tool for infants, which has the potential to support the diagnosis of a range of neurological disorders, including cerebral palsy.

1 Introduction

Infant pose estimation has a number of practical applications. The primary motivator for this work is its role in the development of an automated General Movement Assessment (GMA). There are a number of disorders, including cerebral palsy, which can be diagnosed via the analysis of an infant's movement quality using a GMA [1]. Currently this is a qualitative process undertaken by trained experts. This impedes the assessment's widespread use due its expense and human variability [2]. Therefore, an automated tool which could be deployed with little training and requiring commonly

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

available equipment could transform the diagnosis of a range of conditions. We see the automation of this process requiring the development of two models namely; an infant pose estimation model, the focus of this work, and a diagnostic classifier. This two model design simplifies the final classification task by the including prior knowledge into the model.

The current state-of-the-art approaches in infant pose estimation are based on random ferns and an empirically adapted deep learning model trained only on adult datasets [3]. This is in contrast to the adult domain in which deep learning convolutional neural networks have represented the state of the art since the introduction of DeepPose by Toshov et al. in 2014 [4]. One reason for this difference is that the adult domain contains large datasets including MPI_INF_3DHP [5] and MPII [6]. However in the infant domain the availability of such datasets is limited. Therefore the training of data hungry CNNs on such small datasets has previously been thought to be infeasible [3].

However this work uses transfer learning to overcome this data shortage. The principle is to pre-train a model on a related task with a large dataset before fine-tuning the same model on the smaller target dataset. Intuitively CNN features that are common between the two tasks do not have to be re-learnt for the target dataset hence boosting performance [7].

2 Methodology

Three distinct deep learning models have been trained, consisting of a bounding box model, a 2D pose estimation model and a 3D lifting network. These form the pipeline shown in Fig 1.

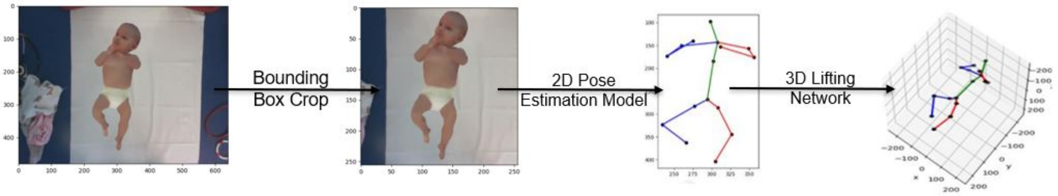

Figure 1: Summary of the three part pose estimation pipeline.

The first step is the cropping of the input image according to the bounding box of the infant which is extracted by a pre-trained Faster R-CNN model [8]. This ensures that the input to the 2D pose estimation model has a subject of consistent scale and also removes extraneous background information.

The next step in the pipeline is 2D pose estimation, the task of estimating the pixel coordinates of a number of keypoints, such as elbows and knees, from an input image. The common approach in the 2D adult domain is to estimate the probability of the keypoint's presence at each pixel location to form a heatmap [9]. To form the ground truths for such a models, a heatmap for each keypoint is produced by placing a symmetrical 2D Gaussian at the keypoint's coordinate location. Inference of the final joint location is found as the coordinate with the maximum value on the predicted heatmap.

The work of Xiao et al. [10] provides a simple baseline model for this heatmap approach and forms the basis of the 2D model architecture proposed here. It is formed of a ResNet50 backbone to which four deconvolutional layers are added which upsample the low level features to produce an output of $\{k \times 64 \times 64\}$ heatmaps where k is the number of keypoints. The model was pre-trained on the MPII adult dataset [6] before fine-tuning on the synthetic infant MINI-RGBD dataset [3].

The final model in the pipeline is the 3D lifting network which takes an input of 2D pixel keypoint locations and outputs the 3D locations of each keypoint relative to the pelvis joint. The lifting model is adapted from the work of Martinez et al. [11]. Simple RELU units are combined with common deep learning techniques such as residual units, batch normalisation and dropout to form a simple but effective deep learning architecture. This model was firstly pre-trained on the MPI_INF_3DHP adult dataset [5] before fine-tuning on the infant MINI-RGBD dataset [3].

To minimise the domain shift between datasets all the keypoint frameworks are mapped to one common definition. This is taken to be the dataset with the fewest keypoints, the MPII dataset [6].

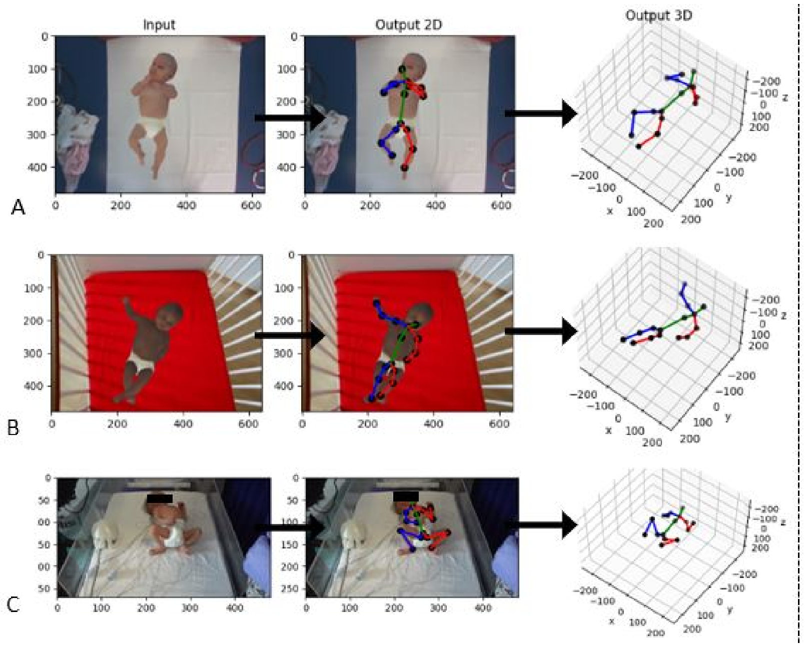

Figure 2: Sample outputs of the model applied to the synthetic MINI-RGBD test set in A and B [3] and real data in C. No 3D ground truth is currently available for the real dataset.

3 Results

Table 1 shows the superior performance of both the 2D and 3D models when compared to the current state-of-the-art results on the synthetic MINI-RGBD dataset [3]. These metrics have been calculated for videos 11 and 12 of the dataset which were held out as a test set during training. Crucially the 3D model proposed here does not require an additional depth channel unlike the current state-of-the-art [12]. Table 1 also shows the increase in accuracy that pre-training the model on adult data has.

Table 1: Comparison of PCKh and AJPE metrics between this projects model, with and without pre-training on adult data. and the current SOTA on videos 11 and 12 of the MINI-RGBD dataset [3].

Model	PCKh 1.0 / %	PCKh 2.0 / %	AJPE / pixel or mm
Ours 2D	93.77	96.05	8.17
Ours 2D w/o Adult Pre-training	82.98	Not given	10.96
Adapted OpenPose (SOTA)	86.54	91.64	Not given
Ours 3D	65.20	88.99	28.47
Ours 3D w/o Adult Pre-training	52.18	84.25	34.18
Hesse et al. (SOTA)	51.09	83.87	44.90

A visualisation of a success and failure case are shown in Fig 2A-B. Fig 2B is a challenging input for the 2D model due to the self occlusion caused by the crossing of the two ankles. The synthetic nature of the data coupled with the lack of clothing makes the distinction between crossing limbs even more difficult. Fig 2B also shows the dependency the 3D lifting network has on the validity of 2D model's output. However, evidence has been seen that the lifting network can reform plausible poses from an incorrect 2D input. A video of the output of the model on the MINI-RGBD test set is included in the supplementary material. This shows the output has a jittery nature. As no temporal information has been used this is not unreasonable and provides potential for future improvement.

The model also has the ability to generalise to real data. This has been tested using the dataset, acquired in the course of the wider research project at Imperial College London, with no further training to the model. The 2D AJPE is 13.82 pixels for this dataset. A video of the outputs on this dataset is included in the supplementary material and a sample output is shown in Fig. 2C

4 Conclusion and Future Research

Future research to refine the model presented here includes the addition of temporal information to the model. However, the largest outstanding research challenge is the generalisation of the model to real data from arbitrary camera. It is reasonable to expect performance gains can be made through fine-tuning the model on a real, non-synthetic dataset. For the 2D model redefining the keypoint framework to the real dataset definition would allow for this. The lack of 3D ground truth makes the addition of ideas from an unsupervised approach such as Chen et al. [13] to the 3D lifting network an interesting avenue for future research.

This paper has presented the first deep learning-based model for 3D infant pose estimation. This has been achieved through the use of transfer learning to adapt models proposed in the adult domain. The model has outperformed the current state-of-the-art 2D and 3D models on the benchmark MINI-RGBD synthetic infant dataset [3]. Furthermore, the 3D model presented here does not require a depth channel unlike the current state-of-the-art random ferns model [12].

5 Broader Impact

The broader impact of this paper is twofold. Firstly the work presented here has shown that deep CNN architectures can be successfully trained on the task of infant pose estimation on a synthetic dataset. Despite the lack of labelled data, we have shown that transfer learning approaches from the adult domain show great promise for future development.

Secondly, the removal of the depth channel requirement opens up possibilities to develop a tool on widely accessible hardware including smartphones, supporting clinicians to diagnose disorders, such as cerebral palsy. Consequently this could lower the average age of diagnosis and thus improve the clinical management and outcomes for infants born with neurological disorders.

Acknowledgements EP/S013687/1

References

[1] Heinz F Prechtl. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early human development*, 1990.

[2] Mijna Hadders-Algra, Annelies MC Mavinkurve-Groothuis, Sabina E Groen, Elisabeth F Stremmelaar, Albert Martijn, and Phillipa R Butcher. Quality of general movements and the development of minor neurological dysfunction at toddler and school age. *Clinical Rehabilitation*, 18(3):287–299, 2004.

[3] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. 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)*, pages 0–0, 2018.

[4] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[5] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[9] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[10] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[11] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[12] Nikolas Hesse, Gregor Stachowiak, Timo Breuer, and Michael Arens. Estimating body pose of infants in depth images using random ferns. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 35–43, 2015.

[13] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019.