Single Image based Infant Body Height and Weight Estimation

Huaijing Shu[1], Lirong Ren[2], Liping Pan[3], Dongmin Huang[1],Hongzhou Lu[3,*], Wenjin Wang[1,*]

[1]Department of Biomedical Engineering, Southern University of Science and Technology, China
[2]Department of Obstetrics, Baoan Hospital of Traditional Chinese Medicine in Shenzhen, China
[3]Department of Neonatal Intensive Care Unit, The Third People's Hospital of Shenzhen, China
[*]Corresponding authors: Hongzhou Lu, Wenjin Wang

Abstract

The collection of infant body data such as height and weight is a useful mean of tracking its growth and wellness. The contact-based measurements using height and weight scales are manual and cumbersome, camera-based methods were proposed to obtain features from face or body for height and weight estimation. In this paper, we created a clinical dataset including 200 newborns collected at obstetrics, and benchmarked four convolutions neural networks for infant height estimation, where MobileNet, VGG16, GoogleNet, and AlexNet were chosen. Moreover, we investigated different MobileNet-based variants for infant weight estimation, including linear regression model, one-task model, and multi-task model. Several sets of experiments were carried out on the newborn dataset to validate the effectiveness of the proposed methods. The results show that the Mean Absolute Error (MAE) of different models are quite similar, with an average MAE < 1.1 *cm and* < 0.28 *kg for height and weight estimation, respectively. Among them, the multi-task MobileNet has better temporal stability given its lower variance of measurement in a video.*

1. Introduction

Height and weight are important anthropometric parameters need to be measured for newborns, i.e. the growth rate of infant weigh can indicate the trend of overweight [1], the low birth weight is closely associated with neonatal mortality [2], the high birth weight may increase the risk of Type-2 diabetes and obesity [3], and height and weight are important factors for osteosarcoma diagnosis [4], etc. Moreover, they are often used for making diagnostic or treatment decisions, i.e. in the emergency department, inpatient department, or consultation clinic [5]. Some of these include the assessment of nutritional status [5] and medication dosage [6]. The height of the infant is typically measured by a height ruler. Two medical staffs need to work together to stabilize the infant's head and feet when measuring the height. The weight data is obtained by a weight scale. The manual measurement is difficult for preterm infants that need to be kept in the incubator to prevent infections [7]. In addition, infants may be suffocated at birth, requiring prompt cardiopulmonary resuscitation and medication. It can pose a safety risk to newborns if the procedure of obtaining the anthropometric parameters is tedious and inefficient [8].

Figure 1. The clinical scenario for infant data collection. The image data were acquired by the smartphone of nurses. The reference data were acquired by the height and weight scales.

Camera-based remote infant monitoring has been studied in the past decade, typically focused on vital signs monitoring such as non-contact optical measurement of heart rate and respiration rate of newborns [9–12]. It triggers our thought if other semantic information like anthropometric parameters (e.g. body height and weight) can be measured by cameras as well in addition to physiological parameters. In this field, almost all related works are about the image-based body height and weight estimation of adults, not on newborns or babies, though this group of the population has a stronger desire on the frequent estimation of height and weight for growth tracking. To estimate the body height, some prior arts rely on the detection of reference length in

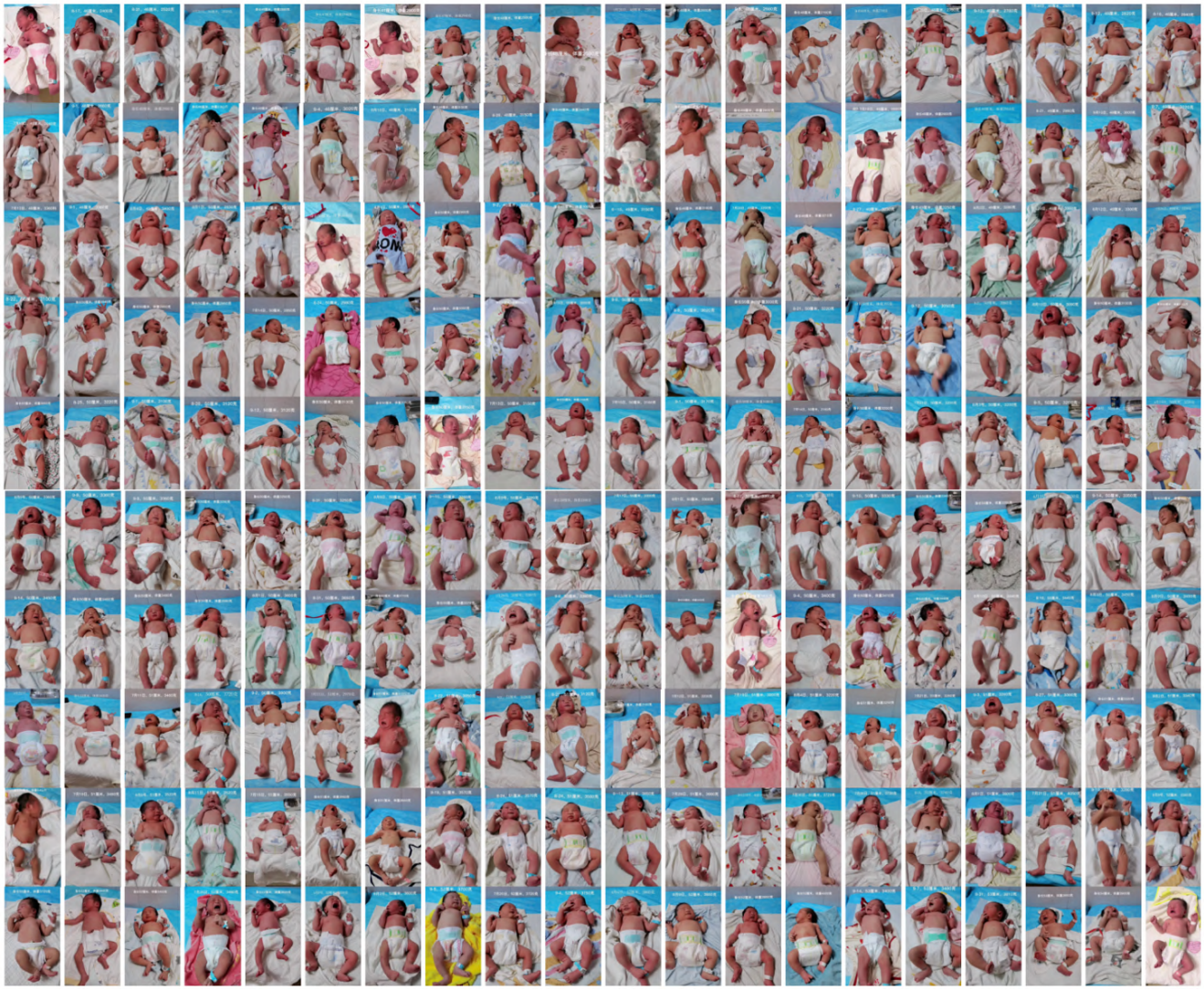


Figure 2. The snapshots of 200 newborns in our created clinical dataset.

images [13, 14]. However, in practice, a reference length may not be available. Some work rely on facial images to estimate height and weight [15, 16], which is convenient for mobile phone applications as it does not require a long distance between the camera and subject to capture the full-body image. However, for newborns, their bodies are relatively small, and in order to prevent infections, it is preferred to keep the measurement at a distance. Thus recording full-body images is more favored for infants. The RGB-D cameras were also explored for 3D body data acquisition [17], where the depth information can provide more accurate information for body size measurement. However, depth cameras are prone to sunlight or the distance of measurement, making it difficult to be used in practical scenarios like in clinics [18]. Therefore, recent advances employed RGB images acquired by a consumer-grade camera for body height and weight estimation [19], which is considered to be more suitable for clinical usage such as deploying the algorithms on the smartphone of nurses.

Inspired by the literature, we conducted a clinical study in obstetrics as shown in Figure 1, and collected a dataset including 200 newborns as shown in Figure 2. We proposed to estimate the infant height and weight from a single 2D body image by the convolutional neural network (CNN). In particular, we benchmarked four different CNNs including VGG16 [20], AlexNet [21], GoogleNet [22], and MobileNet V2 [23], and chose MobileNet V2 as the backbone for height estimation due to its efficiency (aimed for mobile phone applications). Background unrelated to the height estimation has been removed from the input image by setting the pixels outside the mask to zero to eliminate the interference. To estimate the weight, we used three different

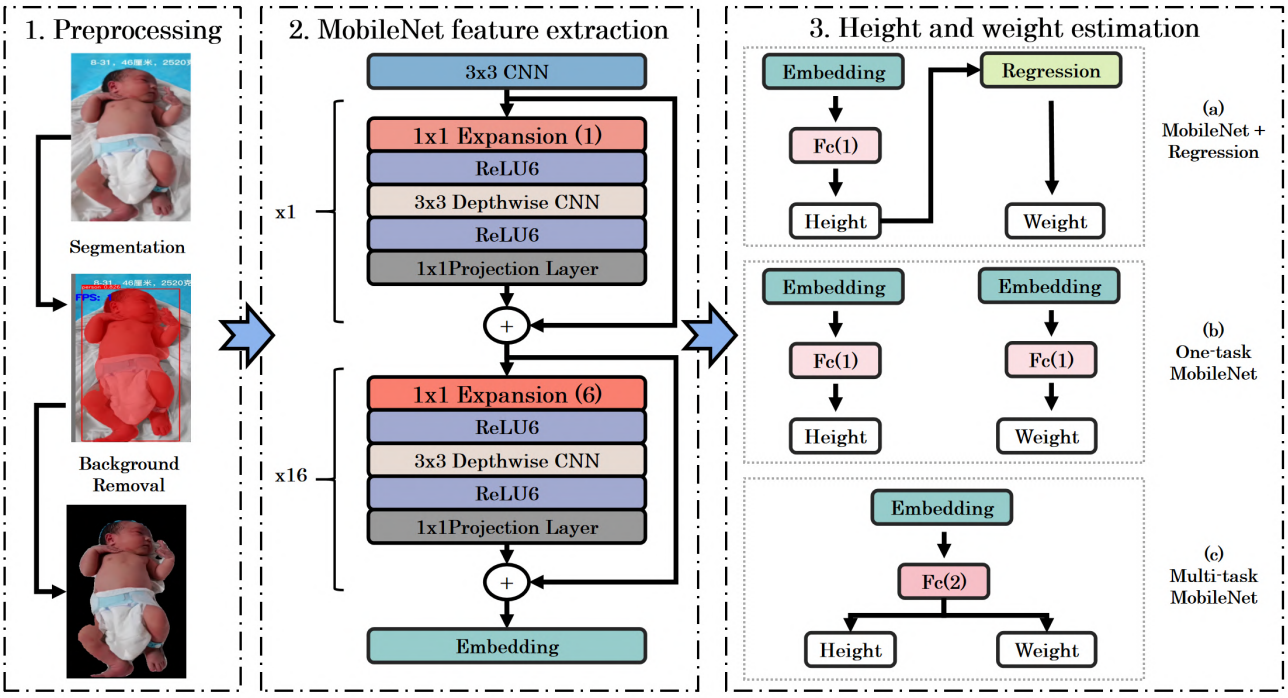


Figure 3. The deep learning models developed for infant height and weight estimation. The system takes a single image as the input and extracts deep features from background-removed version using MobileNet for height estimation. The weight is estimated by three different models following the height estimation.

strategies as shown in Figure 3, i.e. use MobileNet to estimate the height and a linear regression model to estimate the weight, use two MobileNets to estimate the weight and height separately, and use a multi-task MobileNet to estimate the weight and height jointly.

To our best knowledge, this is the first clinical study that exploits the mobile phone camera for infant height and weight estimation. The main contributions of this work are as follows: (i) a clinical dataset was created to facilitate the research on camera-based infant body weight and height estimation; (ii) three CNN-based frameworks were developed to investigate the feasibility of this application. The benchmark results show that three MobileNet-based frameworks have similar MAE for height and weight estimation with an average MAE smaller than 1.1 cm and 0.28 kg for height and weight estimation, respectively. The multi-task MobileNet shows better temporal stability of video-based measurement in the benchmark.

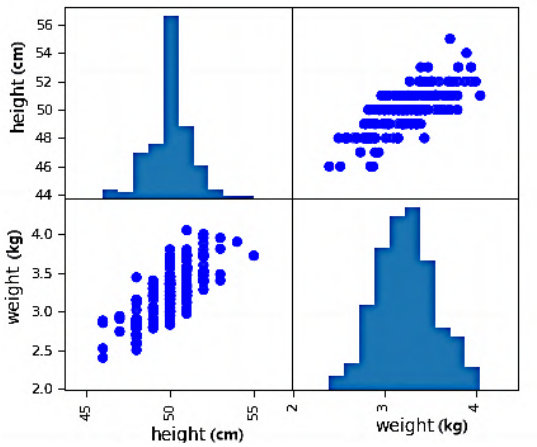
2. Materials and methods

2.1. Clinical infant dataset

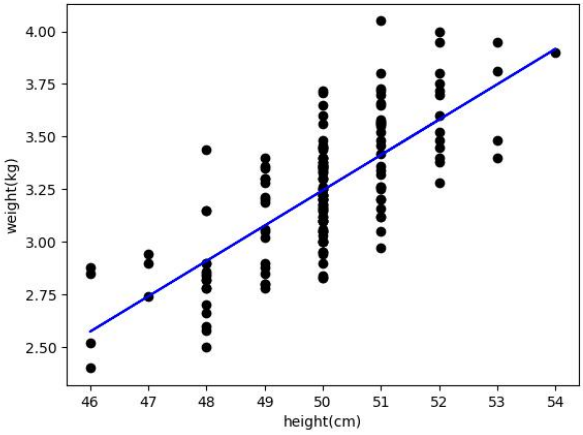
The main focus of this paper is to develop methods for non-contact growth sign monitoring of infants. We first describe the clinical trial that created the benchmark dataset, and then the network architectures for infant body height and weight estimation.

The clinical study involves 200 newborns with < 1 hour of birth (see Figure 2). The study was approved by the hospital Institutional Review Board (IRB) and written consent forms were obtained from the parents of infants. The videos were recorded by the nurse of obstetrics in the Baoan Hospital of Traditional Chinese Medicine in Shenzhen, China, using the 48MP rear camera of a mobile phone in an unconstrained environment. To ensure adequate data sampling, the video length of each newborn was between 20 and 35 seconds, with an average recording length of 30 seconds. The reference height and weight values of each infant were manually obtained by height and weight scales. The averaged height and weight of 200 newborns are 50.0 cm and 3.5 kg, respectively (see their distributions in Figure 4).

As the pre-processing of the dataset, the videos were split into frames and resampled to 224×224 pixels (as the input of CNN network) with the aspect ratio unchanged. As the image background is irrelevant to the estimation of body height and weight, it was segmented by YOLO-v7 [24] in combination with Detectron2 [25], which is a flexible and efficient approach for instance segmentation, and removed before network training, i.e. the pixels outside the mask contour are set to zero to attenuate the background.



(a) The correlation and distribution of 200 infants' height and weight.



(b) The linear regression between 200 infants' height and weight.

Figure 4. The analysis of benchmark dataset and the modeling between the infant's height and weight.

2.2. Network architecture

Based on the background-removed infant body images, three different network models were built to estimate height and weight (see Figure 3): (i) a one-task MobileNet to predict height and a separate linear regression model to estimate weight; (ii) two one-task MobileNets to predict height and weight separately; and (iii) a multi-task MobileNet to predict height and weight simultaneously. Three models are described in detail in the following text.

Height estimation based on MobileNet and weight estimation based on linear regression: We applied the MobileNet V2, an effective and lightweight network, as the backbone to extract features for height estimation. As shown in Figure 3, it is composed of multiple *bottelneck residual block* in series, and each *bottelneck residual block* consists of three important layers: (i) *expansion layer*. It expands the low-dimensional features to high-dimensional features using *pointwise convolution* to enhance the underlying representation; (ii) *depthwise convolution*. It used *depthwise convolution* to calculate the linear combination of input channels to construct new features as deep features; (iii) *projection layer*. It maps the features to the low-dimensional space using *pointwise convolution*, and uses linear transformation to replace ReLU calculation to eliminate the loss of information. We fine-tune the original network by replacing the last 1000 output fully-connected layer with the single output, for estimating the height. Since the measurement of weight is less straightforward than that of height in an image, we analyzed the relationship between height and weight (see Figure 4a). The scatter plot shows that the correlation between the height and weight of infants is somewhat linear, which can be modeled using Pearson Correlation. We established a simple linear regression equation in between as follows:

$$W = \alpha \cdot H + \beta, \tag{1}$$

where H and W denote height and weight respectively; α and β are model parameters to be estimated by least-squares regression.

Height and weight estimation based on one-task MobileNet separately: We developed two models based on MobileNet V2 to estimate height and weight separately. The backbone used the same structure as the first model (MobileNet). To estimate the weight and height, we fine-tuned the network in two separate models by using 1 output instead of 1000 (Figure 3b) and we used separate MSE loss function for each task.

Height and weight estimation based on multi-task network: We established a multi-task model based on MobileNet V2 to estimate height and weight jointly. In the model, we kept the original backbone and modified the last layer with 2 outputs instead of 1000. One output is for height estimation and the other is for weight estimation (Figure 3c). In addition, we used separate MSE to calculate the loss of height and weight. The loss of the model, backpropagated to the input layer, is calculated by the sum of height loss and weight loss. Therefore, we could balance the learning between height and weight to avoid the case that either task dominates the network weights [26].

3. Experiments and results

In this section, we evaluated the proposed methods on the collected newborn dataset. The benchmark dataset includes 200 infants with 800 images for each on average.

3.1. Implementation details

Network training procedures: In the MobileNet-based deep feature extraction, the images need to be resized to feed into the network. We downscaled the whole image diagonally and ensure the long side to be 224 pixels. Then, we set the length of the short edge to 224 pixels by padding with zeros. This generates an image with 224×224 pixels without the change of aspect ratio. Referring to ImageNet [27], we normalize the image tensor with mean values of [0.485, 0.456, 0.406] and standard values [0.229, 0.224, 0.225] when sending it to the network. The learning rate is kept the same as [28] (0.0001) and the weight decay is set to 1e-3. The MobileNet V2 was pre-trained on ImageNet and fine-tune using the training data. All models were evaluated using 5-fold cross-validation on our newborn dataset. Here, the data from one infant only exist in either the training set or the testing set. Depending on the model structure, the output fully-connected layer is modified to either 1 or 2 outputs. All experiments were done by Pytorch [29], the deep learning platform is implemented on NVIDIA GeForce RTX 3050Ti Laptop GPU.

Network evaluation metrics: We adopted Mean Square Error (MSE) as the loss function:

$$MSE = \frac{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}{N}, \tag{2}$$

where $\bar{y}_i$ is the ground-truth value of infant weight or height in the i-th image; y_i is the estimated value for the i-th image; N is the total number of training images. The criteria for evaluating the quality of results are Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}_i|, \tag{3}$$

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \bar{y}_i}{y_i}\right|. \tag{4}$$

The larger values of MAE and MAPE mean larger errors of predicted values. When the prediction perfectly matches the reference, both are 0.

Linear equation fitting: Each fold creates a linear equation based on the current fold training data. The regression model was figured out by scikit-learn [30]. The coefficient of the regression model trained with 5-fold training data are 0.1680, 0.1717, 0.1639, 0.1635, and 0.1483 respectively. The intercept of the regression model trained with 5-fold training data are -5.1577, -5.3349, -4.9524, -4.9461, and -4.1867, respectively. The example of regression line of the equation is shown in Figure 4b.

3.2. Performance and discussion

Comparison of networks for feature extraction: Table 1 shows the quantitative comparison between the used MobileNet and the other three networks: VGG16, GoogleNet, and AlexNet. The results show that the performance of MobileNet is similar to other networks, but it has the least number of parameters, which is a lightweight and efficient option preferred for embedded implementation on mobile phones in the future work.

Table 1. MAE, MAPE(%), and number of parameters of different CNNs for height estimation using 5-fold validation.

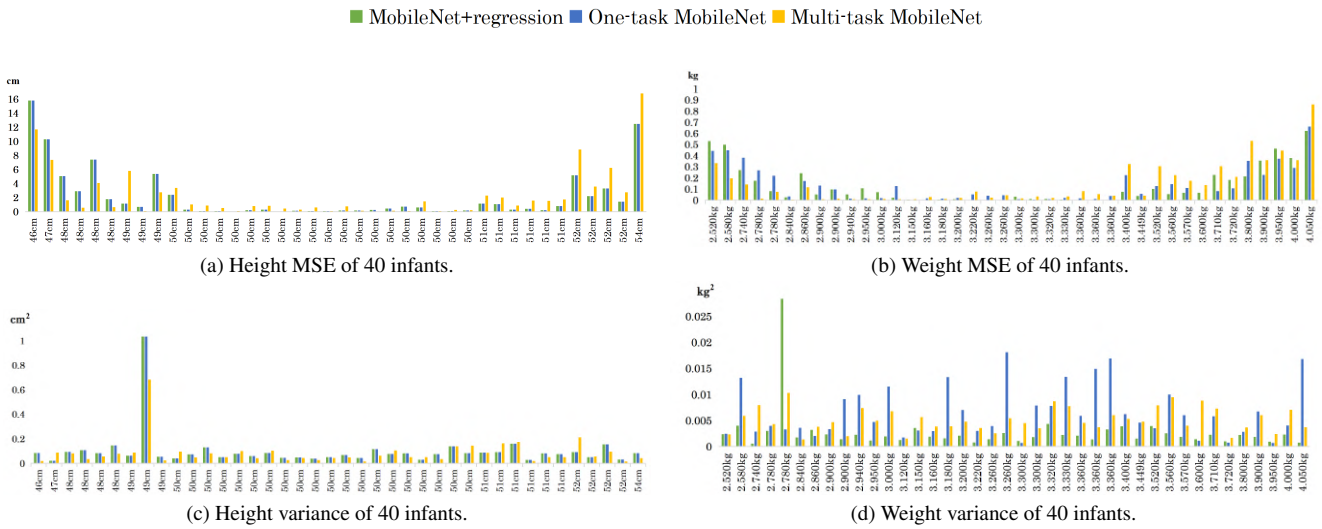
Method	MAE (cm)	MAPE (%)	Param (M)
VGG16 [20]	1.008	2.128	14.74
GoogleNet [22]	1.021	2.418	5.60
AlexNet [21]	1.063	2.260	2.48
MobileNet [23]	1.068	2.440	**2.23**

Background removal: To verify if background removal has a positive effect on height and weight estimation, we compared the MobileNet-based methods using images of 40 testing infants with and without background. The results show that using the same model, the MSE obtained on images without background are clearly reduced, i.e. for MobileNet with linear regression, the weight MSE reduced 20 g and the height MSE reduced 1.1 cm as compared with the original images; for one-task MobileNet the height MSE reduced 1.1 cm and weight MSE reduced 1 g, and for multi-task MobileNet the weight MSE reduced 120 g and the height MSE reduced 2.3 cm. The improvement of background removal is somehow expected and reasonable as it can eliminate the interference that is not relevant to the estimation of body parameters. Some concrete examples on the comparison of images with and without background can be found in Figure 7.

Comparison of different models for height and weight estimation: We compared three types of models as shown in Figure 3. Table 2 shows that the three models have rather similar performance in terms of MAE and MAPE, i.e. MAE is $<$ 1.1 cm and MAPE is $<$ 2.2% for height estimation, and MAE is $<$ 0.28 kg and MAPE is $<$ 9% for weight estimation. The average MSE for 40 testing infants is shown in Figure 5a and Figure 5b. Both show that infants with lower and higher values of height and weight have relatively poor results, which means that our network has worse performance on underweight and overweight infants due to the imbalanced distribution of training data. Since an infant height and weight will not be changed in a video though the prediction may vary, we analyzed the temporal stability of three models using 40 infant videos, where the temporal variation of prediction results given by each video frame is used to quantify the stability. Figure 5c and Figure 5d show that the multi-task model has the best temporal stability as its variances are lowest in most cases, which may be attributed to the mutual constraints between

Table 2. MAE and MAPE (%) of height and weight obtained by different models in 5-fold cross-validation.

Model	Height MAE (cm)	Height MAPE (%)	Weight MAE (kg)	Weight MAPE (%)
MobileNet + Regression	1.041	2.086	0.239	7.518
One-task MobileNet	1.041	2.086	0.252	8.883
Multi-task MobileNet	1.086	2.176	0.233	7.261



(a) Height MSE of 40 infants.

(b) Weight MSE of 40 infants.

(c) Height variance of 40 infants.

(d) Weight variance of 40 infants.

Figure 5. MSE and variance of 40 testing infants processed by three different models. Infants with less training data distributed on either side perform relatively poor. The multi-task MobileNet seems more stable in video-based estimation, as the relationship between height and weight is exploited by the network to stabilize the prediction.

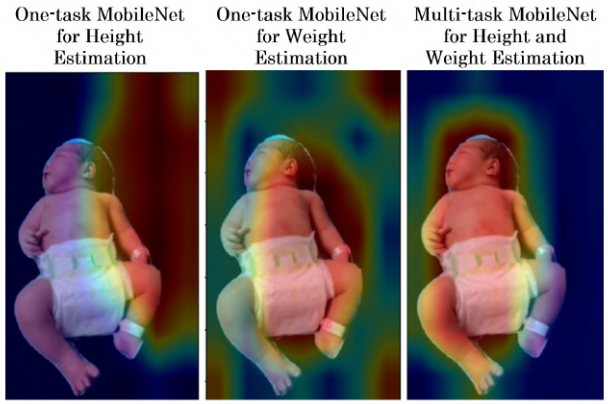


Figure 6. The intermediate activation heatmaps of three models, where the redder areas denote the image parts that have higher contributions to the estimation.

height and weight in the process of loss minimization that makes the overall prediction more stable. To further analyze this phenomenon, we visualize the intermediate activation heatmaps of three methods in Figure 6, which can somehow explain the better stability of the multi-task model. However, this is also counted for large errors in the multi-task network in specific cases, as the prediction errors produced by height or weight can affect each other.

Challenging factors for this application: In the benchmark, we found that some infants performed poorly on height prediction (height MAE $>$ 5.0 cm) regardless the chosen model. We compared images of infants with high-quality performance (height MAE $<$ 3.5 cm) and images of infants with low-quality performance (height MAE $>$ 5.0 cm) at the same height, and compared infant images with high-quality and low-quality performance under the similar pose. The concrete examples are shown in Figure 8. Our observations are follows: (i) camera tilting angle may lead to the degradation in height and weight estimation as the viewing angle is changed. We found that a large portion of poor results are taken in a large tilted viewing-angle; (ii) it is possible that the size of diaper may influence the estimation of height and weight. We see that some poor measurements are from the infants with longer diapers; (iii) in some of examples showing dropped quality, we found that the infant body was not fully captured by the camera. The incomplete data could be a reason for poor results; (iv) background removal has a positive effect on the estimation. But the segmentation errors may influence the estimation. The

	With background examples					Background removal examples				
Test image										
Reference H/W	48cm / 2.780kg	50cm / 2.950kg	50cm / 2.840kg	52cm /3.600kg	54cm / 3.900kg	48cm / 2.780kg	50cm / 2.950kg	50cm / 2.840kg	52cm /3.600kg	54cm / 3.900kg
MobileNet+ Regression H/W MSE	1.69cm / 0.13kg	1.87cm / 0.00kg	1.90cm / 0.03kg	4.79cm / 0.15kg	17.28cm / 0.48kg	1.79cm / 0.12kg	0.12cm / 0.12kg	1.29cm / 0.04kg	1.81cm / 0.06kg	12.46cm / 0.35kg
One-task MobileNet H/W MSE	1.69cm / 0.28kg	1.87cm / 0.09kg	1.90cm / 0.09kg	4.79cm / 0.04kg	17.28cm / 0.29kg	1.79cm / 0.28kg	0.12cm / 0.00kg	1.29cm / 0.06kg	1.81cm / 0.00kg	12.46cm / 0.23kg
Multi-task MobileNet H/W MSE	1.74cm / 0.02kg	1.54cm / 0.01kg	0.96cm / 0.00kg	3.52cm / 0.10kg	17.27cm / 0.49kg	0.72cm / 0.01kg	0.40cm / 0.00kg	2.68cm / 0.00kg	5.53 cm / 0.18kg	16.79cm / 0.36kg

Figure 7. MSE of height and weight estimation from the images with and without background. Three MobileNet-based models are compared. The darker the color, the better the result.

	Low-quality detection examples					High-quality detection examples				
Test image										
Reference H/W	46cm / 2.520kg	54cm / 3.900kg	48cm / 2.860kg	49cm / 3.300kg	52cm / 3.520kg	48cm / 2.780kg	48cm / 2.900kg	46cm / 2.400kg	49cm / 2.900kg	52cm / 4.000kg
MobileNet+ Regression H/W MSE	15.77cm / 0.53kg	12.46cm / 0.35kg	7.40cm / 0.24kg	5.34cm / 0.03kg	5.15cm / 0.10kg	1.79cm / 0.12kg	1.78cm / 0.10kg	3.37cm / 0.23kg	0.66cm / 0.10kg	1.41cm / 0.38kg
One-task MobileNet H/W MSE	15.77cm / 0.44kg	12.46cm / 0.23kg	7.40cm / 0.17kg	5.34cm / 0.01kg	5.15cm / 0.13kg	1.79cm / 0.28kg	1.78cm / 0.10kg	3.37cm / 0.26kg	0.66cm / 0.10kg	1.41cm / 0.29kg
Multi-task MobileNet H/W MSE	11.68cm / 0.33kg	16.79cm / 0.36kg	4.09cm / 0.12kg	2.73cm /0.01kg	8.83cm / 0.30kg	0.72cm / 0.01kg	0.61cm / 0.01kg	1.23cm / 0.05kg	0.03cm / 0.01kg	2.74cm / 0.36kg

Figure 8. MSE of height and weight estimation from low-quality (height MAE $>$ 5.0 cm) and high-quality (height MAE $<$ 3.5 cm) samples. Three MobileNet-based models are compared. The darker the color, the better the result.

images with clutter background that cannot be completely removed by foreground segmentation may have larger errors in height and weight estimation.

4. Conclusions

In this work, we created a clinical dataset with 200 newborns to validate the concept of using a single image for infant body height and weight estimation. Based on the background-removed body images, we investigated different CNN structures and chosen MobileNet as the backbone for deep feature extraction to estimate the height. Three types of models were developed for weight estimation, including linear regression model, one-task MobileNet, and multi-task MobileNet. The benchmark shows that the MAE of different models are similar, which are in a decent error range, i.e. average MAE for height estimation is < 1.1 cm and average MAE for weight estimation is < 0.28 kg. In particular, multi-task MobileNet shows better temporal stability. In our future work, we may add attention modules to the network to force CNN to focus on the image parts that are more relevant to the measurement of body parameters and deploy the verified model on mobile phones.

Acknowledgements

This work is supported by the National Key R&D Program of China (2022YFC2407800), General Program of National Natural Science Foundation of China (62271241), Guangdong Basic and Applied Basic Research Foundation (2023A1515012983), and Shenzhen Fundamental Research Program (JCYJ20220530112601003).

References

[1] Dennison, Barbara A, Edmunds, and et al. Rapid infant weight gain predicts childhood overweight. *Obesity*, 14(3):491–499, 2006. 1

[2] Paneth N. S. The problem of low birth weight. *The Future of children*, 5:19–34, 1995. 1

[3] Johnsson, Inger Wahlström, Haglund, and et al. A high birth weight is associated with increased risk of type 2 diabetes and obesity. *Pediatric obesity*, 10(2):77–83, 2015. 1

[4] Mirabello, Lisa, Pfeiffer, and et al. Height at diagnosis and birth-weight as risk factors for osteosarcoma. *Cancer Causes & Control*, 22:899–908, 2011. 1

[5] Waterlow, John Conrad, Buzina, and et al. The presentation and use of height and weight data for comparing the nutritional status of groups of children under the age of 10 years. *Bulletin of the world Health Organization*, 55(4):489, 1977. 1

[6] Lubitz, Deborah S, Seidel, and et al. A rapid method for estimating weight and resuscitation drug dosages from length in the pediatric age group. *Annals of emergency medicine*, 17(6):576–581, 1988. 1

[7] Baker and Jeffrey P. The incubator and the medical discovery of the premature infant. *Journal of Perinatology*, 20(5):321–328, 2000. 1

[8] Eke, CB, Ubesie, and et al. Comparison of actual (measured) weights and heights with the standard formula methods of estimation among children in enugu. *Nigerian Journal of Paediatrics*, 41(4):307–311, 2014. 1

[9] Zeng Yongshen, Song Xiaoyan, Chen, and et al. A multimodal clinical dataset for critically-ill and premature infant monitoring: Eeg and videos. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5, 2022. 1

[10] Klaessens, John H G M, Marlies Born, Veen, and et al. Development of a baby friendly non-contact method for measuring vital signs: first results of clinical measurements in an open incubator at a neonatal intensive care unit. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, 8935, 01 2014. 1

[11] Abbas K. Abbas, Konrad Heimann, Katrin Jergus, and et al. Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *BioMedical Engineering OnLine*, 10:93 – 93, 2011. 1

[12] Abbas K. Abbas, Id, and et al. Review of biomedical applications of contactless imaging of neonates using infrared thermography and beyond, 08 2018. 1

[13] N. Saitoh, K. Kurosawa, and K. Kuroki. Study on height measurement from a single view. In *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, volume 3, pages 523–526 vol.3, 1999. 2

[14] Criminisi, Antonio, Zisserman, and et al. New approach to obtain height measurements from video. In *Investigation and Forensic Science Technologies*, volume 3576, pages 227–238. SPIE, 1999. 2

[15] Kocabey, Enes, Camurcu, and et al. Face-to-bmi: Using computer vision to infer body mass index on social media. In *Eleventh international AAAI conference on web and social media*, 2017. 2

[16] Dantcheva, Antitza, Bremond, and et al. Show me your face and i will tell you your height, weight and body mass index. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3555–3560. IEEE, 2018. 2

[17] Nguyen, Tam V, Feng, and et al. Seeing human weight from a single rgb-d image. *Journal of Computer Science and Technology*, 29(5):777–784, 2014. 2

[18] Jun-Ming Lu and Mao-Jiun J. Wang. Automated anthropometric data collection using 3d whole body scanners. *Expert Systems with Applications*, 35(1):407–414, 2008. 2

[19] Altinigne, Can Yilmaz, Thanou, and et al. Height and weight estimation from unconstrained images. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2298–2302, 2020. 2

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015. 2, 5

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 2, 5

[22] Christian Szegedy, Wei Liu, Yangqing Jia, and et al. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 2, 5

[23] Sandler, Mark, Howard, and et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 2, 5

[24] Wang, Chien-Yao, Bochkovskiy, and et al. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 07 2022. 3

[25] Yuxin Wu, Alexander Kirillov, Francisco Massa, and et al. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[26] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, and et al. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:3614–3633, 2020. 4

[27] Deng, Jia, Dong, and et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[29] Adam Paszke, Sam Gross, Francisco Massa, and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 5

[30] Fabian Pedregosa, Gael Varoquaux, Gramfort, and et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012. 5