

NurtureNet: A Multi-task Video-based Approach for Newborn Anthropometry

SUPPLEMENTARY MATERIAL

Yash Khandelwal^{1†} Mayur Arvind¹ Sriram Kumar¹ Ashish Gupta¹
Sachin Kumar Danisetty¹ Piyush Bagad² Anish Madan² Mayank Lunayach²
Aditya Annavajjala² Abhishek Maiti² Sansiddh Jain² Aman Dalmia²
Namrata Deka² Jerome White² Jigar Doshi² Angjoo Kanazawa³
Rahul Panicker² Alpan Raval¹ Srinivas Rana¹ Makarand Tapaswi^{1†}

Wadhvani Institute for Artificial Intelligence (WIAI)

¹currently at WIAI, ²work done while at WIAI; ³UC Berkeley

† {yash, makarand}@wadhvaniai.org

Appendix

We present additional details with regard to the data collection procedure (Sec. A) and the data validation process to ensure that the models see correct inputs (Sec. B). Next, we present details related to the experiments: Sec. C discusses the baseline approach, while Sec. D presents clarification regarding metrics and further analysis.

A. Data Collection Process

As described in the main paper, each baby is visited multiple times in the first 6 weeks of life. The data collector visits and captures videos of the baby around the 3, 7, 14, 21, 28, and 42 days after birth to match the health program’s recommended schedule. However, due to field and logistical challenges, we do encourage the data collector to visit the newborn within a ± 2 day window. This gives us an average of 3.75 visits per newborn.

The data collectors are trained to capture a video by starting from the top of the baby and making a smooth arc as illustrated in Fig. 1.

Enrolment. At the first visit, the baby is enrolled using a custom-developed mobile application to ensure data security. The application generates automatic reminders for the data collector to do follow-up visits. Prior to enrolment, the data collectors explain the project to the parents and obtain their informed consent in the local language. During enrolment, we capture basic information such as the mother’s and newborn’s name, address, sex, mode of delivery, date of birth, and weight at birth.

At each visit the data collectors are trained to adhere to the following protocol:

1. After greeting the parents, the first task is to setup the

video capture environment: find a flat, well-lit area in the house, arrange for a bedsheet on which the baby will be placed, and prepare the reference objects.

- Next, the digital weighing machine is prepared for measuring ground-truth. The baby is brought in and its clothes are removed. The newborn is successively placed three times on the weighing machine and readings are noted for each measurement. The whole process is captured in a video to ensure adherence to protocol (see Sec. B.3). As indicated in the main paper, we ensure high quality ground-truth (10 g least count) by using a custom-built, calibrated, and certified weighing machine that averages weight over time.
- We then capture three videos of the baby with different reference object conditions: no reference object, chessboard (♔), and the wooden ruler (📏). For each video, the data collector places the appropriate reference object and makes an arc around the baby as indicated in Fig. 2 of the main paper. We attempt to capture the newborn’s shape by making a steady arc around it while ensuring minimal motion blur (due to camera motion) and that the newborn and the reference object are in the field of view at all times.
- The data collector also measures the newborn’s length using an infantometer, and its head and chest circumference using tape measures. We train our models in a multi-task manner to predict these measurements.
- Finally, an oral health assessment is performed by quizzing the parents on aspects such as feeding status, breathing rate, appearance, muscle tone, and discharge from the eyes or umbilicus.
- In case of any concerns or anomalous responses, the data collectors counsel the parents on potential recourses to

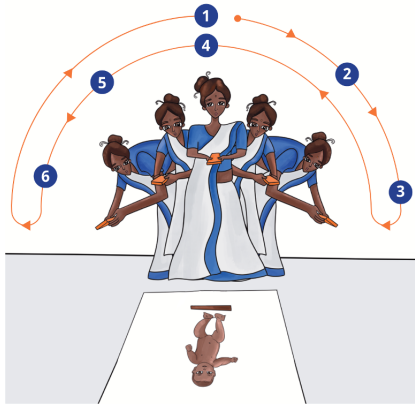


Figure 1. Video recording process followed by health workers to capture the newborn from multiple viewing angles.

address them.

The data is automatically synced to secure cloud storage when the mobile device has access to the internet (note that rural areas where data collection happens may not necessarily have access to the internet) and de-identified before sharing for further processing.

B. Data Validation Criteria

We are interested in understanding the data quality through various annotations related to the environment, the use of appropriate reference object, clothing artifacts on the newborn, and ground-truth. We obtained videos of 16,612 visits across two geographically diverse regions. A team of 5 annotators was trained on the prescribed protocol, and 2 annotators independently annotated each video. After validation, we were left with 12,901 usable visits. Table 1 enlists the criteria used to discard visits. This validation protocol involves three sequential steps as described in the subsections below.

B.1. Environment Validation

Our data is collected in everyday houses in rural, low resource areas in low- and middle-income countries (LMICs) where the video capture environment is unconstrained. This leads to diverse variations in the visual settings across the captured videos and props up classic vision challenges related to poor lighting; bedsheets of different colors, shapes, and textures; and other challenges related to data collection, such as the lack of a video capture setup potentially leading to motion blur and inconsistency in recorded videos. This is far from clinical settings (*e.g.* a hospital) where all newborns may be brought to the same room or even the same bed and captured by the same data collector (nurse) with the same device, making the vision problem easier to solve.



Our first check ensures that each video has a newborn with the correct reference object. Note that for each visit

Criteria	# discarded visits	% discarded
Environment Validation	53	0.3%
Video Quality Validation	441	2.6%
Weight Validation	3200	19.2%
<40 frames in video	17	0.1%
Total	3711	22.2%

Table 1. Number of visits discarded based on all the data validation criteria.

Criteria	# discarded visits
Newborn is not visible	20
Newborn is wearing clothes	47
Readings beyond 50 g of each other	982
Other problems	2151
Total	3200

Table 2. Number of visits discarded due to failure in one or more weight validation criteria. We apply rules strictly to ensure high quality and accurate ground-truth, both for training and evaluation. See Sec. B.3 for a detailed explanation.

we collect three videos with different reference object conditions: no reference object, with a , and with a . Due to the simple nature of this task, we use unanimity to ensure that the annotations are correct. We remove 53 visits after this check leaving us with 16,559 visits that are passed on to the next stage.

B.2. Video Quality Validation

We validate the quality of the videos with the aid of a questionnaire to determine the quality of data collection and ensure adherence to protocol. The questions are: (i) is the newborn wearing clothes? (ii) is the newborn cropped? (iii) is the reference object cropped? (iv) is there good and sufficient light? (v) is the video blurry? (vi) is the newborn and reference object on the same plane? (vii) are there other humans visible in the video? (viii) is the arc smooth or jerky? and (ix) is the newborn captured well from both left and right side angles (*i.e.* how complete is the arc)?

We accept partial failures (*e.g.* newborn cropped for 1 to 3 s) in most of the above criteria and observe that complete failures (correspondingly, newborn cropped for ≥ 3 s) are quite rare. We plan to use the annotations for future analysis and potential studies in error attribution. As we are interested in building a robust anthropometry estimation system, we realize that all videos will not be captured well during deployment. We discard 441 visits in this process and are left with 16,118 visits.

B.3. Ground-truth Weight Validation

The third and final validation check concerns the ground-truth weight. It involves annotators watching the video

Representation	Feature dimensionality
Hu moments	350
Regionprops	300
HOG	7200

Table 3. Hand-crafted feature dimensionality across 25 frames.

recording in which the ground-truth weight of the newborn is captured and recording the observed weight. Recall that the newborn is placed thrice on the weighing machine leading to a total of 6 weight readings across two annotators.

Visits that have 4 of 6 weight readings in agreement are directly accepted. Alternatively, if no reading is more than 50 g away from the mean of all 6 readings, we accept the visit. All other visits are passed through the criteria below. We discard the visits if any of the following is true: (i) the newborn is not visible on the weighing machine; (ii) newborn is wearing clothes while being placed on the machine; (iii) readings are not stable with two or more than two readings varying beyond 50 g; and (iv) a large chunk is attributed to other problems such as the weighing machine that may not be placed on a proper flat surface or is not visible in the video due to occlusions or lack of focus or glare, someone’s hand touching the weighing pan, the newborn’s limbs are touching a nearby wall, *etc.* Visits that do not fail any of the 4 rejection criteria are also accepted. Table 2 shows the counts of the rejection criteria where we discard the visits.

The stringent ground-truth weight annotation protocol along with our weighing machine with a hold function allows us to capture highly accurate values of the ground-truth weight. We removed 3200 visits in this process and are left with 12,918 visits. 17 more visits are finally removed since they have short videos (less than 40 frames as required for subsampling). Finally, 12,901 videos are used as part of our experiments.

C. Baselines

For all our baseline experiments, we use `scikit-image` [4] for extracting the hand-crafted features. Hu moments and Regionprops are extracted from the binary masks of the baby and wooden ruler regions, while the HOG features are extracted from the combined baby and wooden ruler regions cropped from the original image. The hand-crafted features across 25 frames for a given video are concatenated together to create the final feature vector (Table 3).

We evaluate three regression models: ordinary least squares based Linear Regression (LR), Multi-Layer Perceptron (MLP), and kernel Support Vector Regressor (SVR). For LR and MLP, we scale the Hu moments with a log transformation to reduce the variability in feature values. The z -

Representation	Feature Scaling	Model	W (g)
HOG	z -score	LR	853.5
HOG	minmax	MLP	578.4
HOG	z -score	SVR	466.7
Hu moments	log	LR	475.1
Hu moments	log	MLP	477.6
Hu moments	z -score	SVR	470.3
Regionprops	z -score	LR	401.9
Regionprops	minmax	MLP	446.5
Regionprops	z -score	SVR	399.4
Hu + Regionprops	log + z -score	LR	398.1
Hu + Regionprops	log + z -score	MLP	529.0
Hu + Regionprops	z -score	SVR	393.0

Table 4. Performance of hand-crafted features on weight estimation with the validation set.

score and minmax feature scalers have been experimented with. For the MLP, we use the `scikit-learn` [2] implementation, and for the kernel SVR, we use the `LIBSVM` [1] implementation. For MLP, we use one hidden layer of 100 units with an initial learning rate of 0.001 and an inverse scaling learning rate scheduler. For SVR that uses the Radial Basis Function (RBF) kernel, the kernel coefficient γ is set to $\frac{1}{d \cdot \Sigma(X)}$, where d is the feature dimensionality and $\Sigma(X)$ is the variance of X . Table 4 shows the performance of hand-crafted features with different models that regress weight. SVR outperforms LR and MLP across all representations with the combined Hu and Regionprops features giving the best performance on weight estimation.

D. Experimental Details and Analysis

We present additional experimental details related to metrics and some analysis.

D.1. Metric: Balanced MAE

The standard metric Mean Absolute Error (MAE) does not take into account the label distribution (*e.g.* majority of the newborns in our dataset have weight between 2.5 to 3.5 kg). As our goal is related to identifying malnutrition in newborns, it is important to get accurate predictions and metrics corresponding to low birth weight newborns. Thus, we use a more equitable and fair metric: Balanced MAE (BMAE), defined as the average of MAE across multiple weight bins. In our experiments, we use bins of 500 g granularity. Based on the weight distribution in the dataset (see Fig. ?? (Left) in the main paper), we set the lower limit to 1 kg and the upper limit to 5.5 kg. The bins are thus defined as follows:

$$B = \{[l, l + 0.5) \mid l \in \{1, 1.5, \dots, 5\}\}, \quad (1)$$

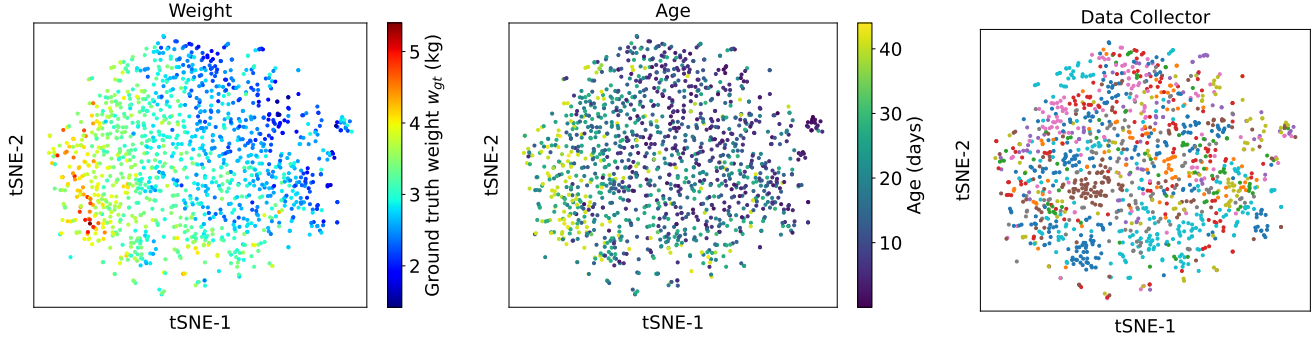


Figure 2. t-SNE embeddings of representations from the video-based model on the validation set. Each dot is colored by different properties: weight (left), age (center), and data collector (right).

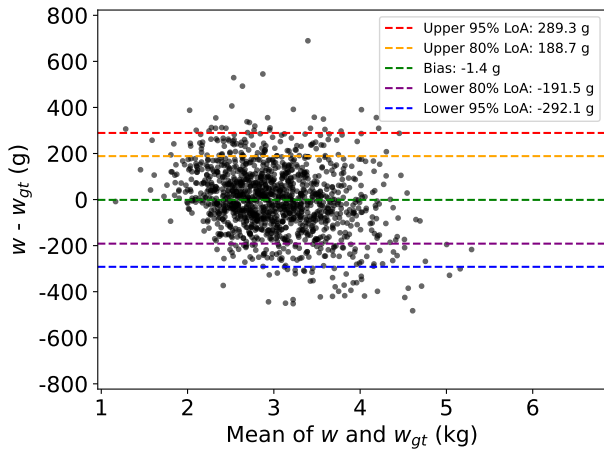


Figure 3. Bland-Altman analysis between ground-truth and NurtureNet’s weight estimates on the test set. Limits of Agreement (LoA) are plotted at 80% and 95% confidence intervals.

and the BMAE metric is defined as:

$$\text{BMAE} = \frac{1}{|B|} \sum_{b \in B} \frac{1}{|b|} \sum_{w_{gt} \in b} |w - w_{gt}|, \quad (2)$$

where $|B|$ is the number of bins and $|b|$ is the number of samples in a particular bin.

D.2. Bland-Altman Plot

We perform a Bland-Altman analysis on the test set to assess the agreement between the predictions of NurtureNet and the ground-truth weight measurements (Fig. 3). The analysis shows negligible bias of -1.4 g indicating a strong agreement of the weight estimates against the ground-truths. Notably the plot largely exhibits homoscedasticity, signifying consistent variability across a significant range of weights. The 80% Limits of Agreement (LoA) is $\sim \pm 190$ g which makes the solution acceptable for deployment based on inputs from public health experts.

D.3. t-SNE plots

Fig. 2 shows t-SNE embeddings [3] for videos from the validation set. We use the simple video-based model for this analysis to visualize the feature space to capture the variation of weight without the influence of multiple tasks or tabular information. (i) In the left plot, colors indicate the true weight of the newborn in kg. We see a smooth color distribution across the embeddings indicating that the model has optimized to a good representation space. (ii) The center plot shows the age of the newborn at the time of data collection in days. We observe a smooth transition here as well. However, there are some higher age babies at the top right and vice versa. (iii) In the right plot, we color the dots by the data collector. A good mix is observed which is desirable to ensure invariance across data collectors.

References

- [1] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011. 3
- [2] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 3
- [3] Laurens van der Maaten and Geoffrey Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008. 4
- [4] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. 3