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

Height Estimation of Children under Five Years using Depth Images

Anusua Trivedi[1], Mohit Jain[1], Nikhil Kumar Gupta[2], Markus Hinsche[2], Prashant Singh[2]
Markus Matiaschek[2], Tristan Behrens[2], Mirco Militeri[1], Cameron Birge[1], Shivangi Kaushik[2]
Archisman Mohapatra[3], Rita Chatterjee[4], Rahul Dodhia[1], Juan Lavista Ferres[1]

Abstract— **Malnutrition is a global health crisis and is a leading cause of death among children under 5 years. Detecting malnutrition requires anthropometric measurements of weight, height, and middle-upper arm circumference. However, measuring them accurately is a challenge, especially in the global south, due to limited resources. In this work, we propose a CNN-based approach to estimate the height of standing children under 5 years from depth images collected using a smartphone. According to the SMART Methodology Manual, the acceptable accuracy for height is less than 1.4 cm. On training our deep learning model on 87131 depth images, our model achieved a mean absolute error of 1.64% on 57064 test images. For 70.3% test images, we estimated height accurately within the acceptable 1.4 cm range. Thus, our proposed solution can accurately detect stunting (low height-for-age) in standing children below 5 years of age.**

I. Introduction

Malnutrition refers to an imbalance of nutrition, both under and over-nutrition. According to the World Health Organization (WHO), malnutrition is a global health crisis and is the reason behind ∼45% of deaths among children under 5 [2]. The risk of malnutrition is especially high among children below the age of 5, and effective early interventions can help overcome the alarming situation. Malnutrition is categorized into undernutrition, micro-nutrient-related malnutrition, and overweight. Deficiency of nutrition, i.e. *undernutrition*, is the leading reason for malnutrition in the global south. It is mainly associated with poor socioeconomic conditions due to unavailability of enough food to eat, infectious diseases, and/or lack of knowledge about young child care. Undernutrition makes the children more vulnerable to other diseases as well, and increases the risk of death. There are three forms of undernutrition: wasting (low weight-for-height), stunting (low height-for-age) and underweight (low weight-for-age).

Measuring malnutrition accurately is challenging. Early detection and intervention require regular anthropometric measurements, measuring weight, height, and middle-upper arm circumference of children under 5. However, such measurements may also have errors due to inexpert data collectors, inadequate tools, and/or poor data management [9]. According to SMART Methodology Manual [5], the acceptable technical error of measurement (*i.e.*, variance between two rounds of height measurement) is less than 1.2 cm, and the acceptable accuracy (*i.e.*, difference between the measured height and ground truth) is less than 1.4 cm.

In this paper, we propose a Convolutional Neural Network (CNN) based method to accurately estimate the height of standing children under 5 from depth images collected using a commercial off-the-shelf smartphone. Overall, we collected data of 3887 children (2581 train data, 1306 test data) aged 2-5 years in rural India. Our approach estimated height with a mean absolute error of 1.64%, and for 70.3% test images, it achieved the acceptable 1.4 cm range. Hence, our solution can detect stunting accurately, by predicting the estimated height with the child's age.

II. Related Work

Estimating anthropometric measurements, especially height, using images has been an active area of research [6], [8], [10]. Researchers have proposed height estimation using a single image [6], images of a subject taken from multiple views [7], [8], and from three-dimensional depth images [10]. A single image of the subject requires a cubical reference object of known dimension along with complex calibration and genetic algorithm [6]. The proposed approach was only evaluated with one subject and achieved a high estimation error of 5.5%. Multiple views (usually 5 or more views) of the subject enables three-dimensional reconstruction of the body surface to estimate height and other anthropometric measurements. Li *et al.* [7] evaluated their multi-view approach with a child mannequin, while Liu *et al.* [8] validated with 31 adults to achieve ∼1% error. More recently, with easy access to depth cameras in Microsoft Kinect and Google Tango devices, depth images have been used for anthropometric measurements. Yin and Zhou [10] used a single depth image and passed it through a four-stage CNN to predict lengths of different body parts and total height. The proposed method was executed on 2136 images collected from 14 adults in 10 different postures. Overall, their method performed well across postures, achieving a total average error of 0.9%, and they found depth images to outperform RGB images. However, in all the above cases, these papers proposed methods to measure height of adults only. To the best of our knowledge, none of the prior approaches were tested with standing children below 5 years, which is the key novelty of our work.

[1] Microsoft {antriv, mohja}@microsoft.com
[2] Child Growth Monitor, Welthungerhilfe
[3] Executive Director at GRID Council, India
[4] (Retired) Professor of Pediatrics, Dr. B C Roy PGI PS, India

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

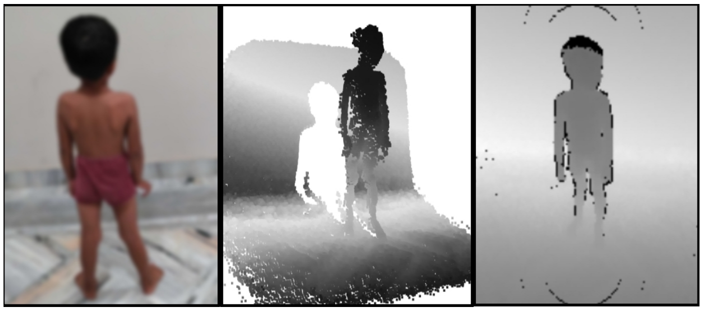

Fig. 1. (from left to right) (a) Back video of a child. (b) Point cloud data. (c) Depth image.

III. Dataset

All the data was collected from two states of India, Rajasthan (Baran district) and Madhya Pradesh (Chatarpur and Sheopur districts), during 2017-2019, using the Child Growth Monitor [1] phone app developed by Welthungerhilfe. We focus on data collected for children who can stand (usually 2-5 years of age) for this work. The data was collected in the regional Anganwadi centers. Anganwadi is a type of rural child care center in India. The data collectors were mostly young adults (20-30 years old) and received a four-day data collection training.

After getting consent from their parents/grandparents, children were asked to stand in front of a solid-colored wall. If needed, a white banner was placed behind the child to replicate a wall. All the videos were recorded using the Lenovo Phab 2 Pro phone, which has a time-of-flight sensor to capture *point cloud data* at 1920x1080 resolution with three frames/second. The point cloud videos were converted into depth images in the data processing stage. For each child, the data collector used the phone app to collect three point cloud videos: (a) *front video*: where the child is facing the camera, (b) *back video*: where the child's back is facing the camera (Figure 1a), and (c) 360-degree video: where the child was asked to spin slowly to capture a 360-degree view of the child. The data collector decided the length of these videos; usually, the front and back videos were 2-4 seconds long, while the 360-degree videos were 5-8 seconds. (Note: For front and back data, a single image would have sufficed, however as children move frequently, we opted for videos.) The data collector ensured that the child's head to toe was fully visible in each video. Next, manual measurements of the ground truth weight, height, and mid-upper arm circumference (MUAC) were taken, using the standardized weight machine, height board, and MUAC tape, respectively. On average, it took 15-20 mins to collect data for a child, involving consent forms, digital videos, and manual measurements. In case the child did not co-operate, they moved to the next child. The child and/or guardian did not receive any incentive for participation.

Overall, data was collected for 3887 children, and the age-wise distribution of the point cloud video dataset is shown in Table I.

Note: The data collection was performed in compliance with the Indian Council of Medical Research 2017 guidelines for biomedical research involving human participants, ensuring the core principles of ethics – autonomy, beneficence, non-maleficence and justice. Participant anonymity, confidentiality and privacy was maintained at all stages.

IV. Data Transformations

Point cloud data is a flexible file format often used to store multidimensional data. In our case, the point cloud data constitutes of a set of points, wherein a single point p represents the three components of three-dimensional space, *i.e.*, x, y, and z values (visualized in Figure 1b and Figure 2). Each collected point cloud video comprises several image frames. As part of data transformation, we extract each point cloud image frame and convert them to depth images, thus obtaining our dataset of 144195 depth images (Table II). Note: As the child is always moving, each image frame is (slightly) different, thus helping us to augment our dataset naturally. The video type—front, back, and 360-degree—based distribution of our depth images is shown in Table II.

A *depth image* is an image that contains information related to the distance of the surfaces of objects from a viewpoint in the real world. For each pixel in a depth image, it has a d distance value ('depth') from the camera sensor coordinate system to the object. Hence, a depth image is represented as a matrix, where each cell (or pixel) contains a single metric depth information (shown in Figure 1c), accessed by two coordinates, u and v (Figure 2).

This transformation from point cloud data to depth image is performed using a projection matrix. The parameters in the projection matrix depend upon the camera specification, including sensor size and focal length. For every point p in the point cloud, there exists a point q in the depth image plane (Figure 2). In other words, the pixel on the depth image plane is a projection of the real world. (Note: Latest phones with depth cameras provide direct access to depth images; hence this data transformation step is not needed.)

As the data was collected for children below the age of 5, it was noisy with several frames in a point cloud video consisting of poor quality data. Hence, the videos were graded manually based on their quality. A video with any frame consisting of noisy data, *i.e.*, no child visible, blurry, too dark, several people present, *etc.*, was rated '*bad*'. Only '*good*' videos were included in the test dataset, resulting

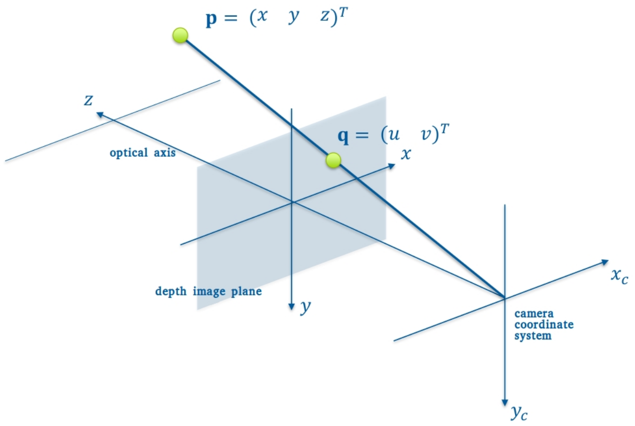

Fig. 2. Point cloud to Depth Image Transformation

TABLE I

AGE-WISE DISTRIBUTION OF TRAINING AND TEST VIDEOS

Age	Total	Training	Test
2-3	1030	712	318
3-4	1370	895	475
4-5	1487	974	513
Total	**3887**	2581	1306

in videos from 1306 children, with 57064 depth images (Table II).

V. METHOD

Our data collector collected three videos of each child using the smartphone, along with manual anthropometric measurements, including the height of the children. We used the point cloud data and manual height labels to train various deep learning models, such as GAPNet and PointNet, focusing on minimizing the mean absolute error in height estimation.

PointNet [3] uses raw point cloud data as input without any emphasis on their ordering, while GAPNet [4] exploits local features by introducing GAPLayer, which assigns different attention weights on the neighborhood for each point. We trained these models on our point cloud data, however even after tuning, the mean absolute error was very high at 4 cm, which was unacceptable. This led us to transform the point cloud data to depth images, which was used to train a Convolutional Neural Network (CNN) based deep learning model for height estimation. Our model consists of 12 convolutional and three dense layers, with padding of 240x180 pixels. We used Rectified Linear Unit (ReLU) as activation function and Mean Squared Error as the loss function.

TABLE II

VIDEO TYPE DISTRIBUTION OF OBTAINED DEPTH IMAGES

Video Type	Total	Training	Test
Front video	38602	24852	13750
Back video	37333	23272	14061
360-degree video	68260	39007	29253
Total	**144195**	87131	57064

TABLE III

SMART MANUAL MEASUREMENT RANGES

	Intra TEM Range	Bias from Supervisor Range
Good	$< 0.4cm$	$< 0.4cm$
Fair	$< 0.6cm$	$< 0.6cm$
Poor	$< 1.2cm$	$< 1.4cm$
Reject	$> 1.2cm$	$> 1.4cm$

TABLE IV

PERFORMANCE OF OUR CNN MODEL ON TEST DATASET

Video Type	In 1.4 cm Range	MAPE
Front	71.77%	1.586%
Back	69.48%	1.671%
360-degree	69.61%	1.680%
Average	**70.28%**	**1.645%**

VI. EVALUATION METRICS

Apart from the mean absolute error and mean absolute percentage error on the test dataset, we also evaluated our approach using the Standardisation Test evaluation metric [5]. A standardisation test for anthropometric measurements is a practical assessment of the data collector's measurement skills. It helps to objectively evaluate the quality—precision, and accuracy—of the measurements taken by each data collector. The standardisation test consists of each collector measuring ten different children twice in two rounds of measurements. Along with the standardisation test for manual measurements, we adapted the standardisation test for our point cloud data. For that, each data collector captured three videos of the child at each round of measurement as well.

For evaluation, we calculated two metrics:

- **Intra TEM (Technical Error of Measurement)** is used to evaluate the measure of *precision* in the standardisation test. For each data collector, it is calculated using the variance between the height predictions on the videos of a child taken in two rounds of measurement.
- **Bias from Supervisor** is used to evaluate the measure of *accuracy* in the standardisation test. For each data collector, it is calculated using the mean of height predictions using the videos of a child compared to the ground truth (*i.e.*, an expert supervisor's manual measurements).

We use guiding principles from the SMART (Standardized Monitoring and Assessment of Relief and Transitions) Methodology Manual [5] to compare the manual versus our models' measurements. Table III shows acceptable limits for Intra TEM and Bias from Supervisor in a standardisation test. According to this table, any Intra TEM prediction with less than 1.2 cm and any Bias from Supervisor prediction with less than 1.4 cm are acceptable.

VII. EVALUATION OF OUR MODEL

Our CNN model achieved a mean absolute error of 1.4 cm and a MAPE (mean absolute percentage error) of 1.64%.

Table IV shows the results on the test dataset using our CNN model, with 71.77% of front videos, 69.48% of back videos, and 69.61% of 360-degree videos were under 1.4 cm acceptable range of height estimation, as per SMART Methodology Manual [5]. We see that the font videos perform the best, even though 360-degree videos offer more surface area. This result is in alignment to the work of Hung et. al. **??**, which shows the front scans are more effective than circumference scans for anthropometric measures. As the test and training set were completely disjoint, it shows that our proposed CNN model works reasonably well on unseen real-world data. Moreover, we collected ten children data (5 male and 5 female, with an average age of 5±1.5 years) using six data collectors and an expert supervisor's ground truth height labels. We use our CNN model to predict the height of the children using the captured front videos (as among the three video types, the front videos performed slightly better) by the data collectors. This standardisation test data was collected during Jan'2020 in Rajasthan, India.

With respect to Intra TEM, comparing manual measurements, all six data collectors performed in the 'Good' range on an average. In contrast, for our CNN model, four data collectors were in the 'Poor' range, and two were in the 'Reject' range.

With respect to Bias for Supervisor, comparing manual measurements, all the data collectors performed in the 'Good' range, and for our CNN model, all data collectors were in the 'Poor' range, thus were acceptable too. This shows that our CNN model achieves accuracy but lacks precision.

VIII. Discussion

One or more forms of malnutrition affect every country in the world. Combating malnutrition in all its forms is one of the most significant challenges for global health. One of the big problems in tackling malnutrition is that it is difficult to identify by conventional means or naked eye whether a child is suffering from malnutrition. Due to flawed data, most of the time, the field-aid workers cannot reach out to children who urgently require assistance. Anthropometric measurements defining malnutrition may not be without errors if taken by inexpert hands and/or inadequate tools. Thus an enhanced ability to improve anthropometric measures in children and determine an individual's propensity to develop or have progressive complications from malnutrition would be of enormous benefit. Recent technological growth has allowed the development of clinical data acquisition and analysis of such data much more straightforward. Efficient and effective smartphone devices will enable us to non-invasively collect more informative images (like point cloud data and depth images) in large populations. In this paper, we show the benefits of combining advanced imaging technology with deep learning methods for estimating height (which can be used to predict stunting) in standing children below 5 years of age. The availability of complex images combined with the rapid evolution of computational data science offers promising opportunities for extracting new inferences and actionable insights that have the potential to improve health outcomes significantly. This more sophisticated data-enriched environment, in turn, has the potential to allow better clinical decision-making through support by automated means, encouraging moves towards intelligent assistance and diagnosis.

IX. Ethics Statement

Welthungerhilfe collected all the patient data for this study. All data is de-identified and anonymized. Consent for the use of this data for research was obtained from the participants. Institute Ethics Committee of All India Institute of Medical Sciences, Patna India (AIIMS IEC) approved our study.

References

[1] Child growth monitor, welthungerhilfe.

[2] M. Blossner, M. De Onis, and A. Pruss-Ustun. Malnutrition: Quantifying the health impact at national and local levels, 2005.

[3] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017.

[4] C. Chen, L. Z. Fragonara, and A. Tsourdos. Gapnet: Graph attention based point neural network for exploiting local feature of point cloud. *CoRR*, abs/1905.08705, 2019.

[5] S. Initiative. Standardized monitoring and assessment for relief and transitions manual 2.0. 2017.

[6] K.-Z. Lee. A simple calibration approach to single view height estimation. In *2012 Ninth Conference on Computer and Robot Vision*, pages 161–166, 2012.

[7] J. Li, M. Sun, H.-C. Chen, Z. Li, and W. Jia. Anthropometric measurements from multi-view images. In *2012 38th Annual Northeast Bioengineering Conference (NEBEC)*, pages 426–427, 2012.

[8] Y. Liu, A. Sowmya, and H. Khamis. Single camera multi-view anthropometric measurement of human height and mid-upper arm circumference using linear regression. *PLOS ONE*, 13(4):e0195600, Apr. 2018.

[9] I. Medhi, M. Jain, A. Tewari, M. Bhavsar, M. Matheke-Fischer, and E. Cutrell. Combating rural child malnutrition through inexpensive mobile phones. In *NordiCHI*, page 635–644. ACM, 2012.

[10] F. Yin and S. Zhou. Accurate estimation of body height from a single depth image via a four-stage developing network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8264–8273, 2020.