

# Accurate Estimation of Body Height from a Single Depth Image via a Four-Stage Developing Network

Fukun Yin\*      Shizhe Zhou\*

College of Computer Science and Electronic Engineering,  
Hunan University, Changsha, China

\*joint first authors: yfk@hnu.edu.cn, shizhe@hnu.edu.cn (corresponding author)

## Abstract

*Non-contact measurement of human body height can be very difficult under some circumstances. In this paper we address the problem of accurately estimating the height of a person with arbitrary postures from a single depth image. By introducing a novel part-based intermediate representation plus a four-stage increasingly complex deep neural network, we manage to achieve significantly higher accuracy than previous methods. We first describe the human body in the form of a segmentation of human torso as four nearly rigid parts and then predict their lengths respectively by 3 CNNs. Instead of directly adding the lengths of these parts together, we further construct another independent developing CNN that combines the intermediate representation, part lengths and depth information together to finally predict the body height results. Here we develop an increasingly complex network architecture and adopt a hybrid pooling to optimize training process. To the best of our knowledge, this is the first method that estimates height only from a single depth image. In experiments our average accuracy reaches at 99.1% for people in various positions and postures.*

## 1. Introduction

In the field of three-dimensional reconstruction, medical treatment, clothes sizing, etc., human height data is indispensable. In most of the cases, we will require the tested person to stand up straight, and then use a meter or other tools to measure the height, which will consume a lot of time and manpower. Especially in actual application, it will be very hard to measure the height if we lack the measuring tools or if the measured person is a child or is injured who cannot stand up straight. Our method can effectively solve these problems as it only needs a single depth image and then outputs reliable results with an average accuracy of 99.1% in milliseconds, saving a lot of manpower and time. More importantly, we do not require the measured person to

stand in a fixed place or a standard posture. They can stand in various postures such as walking, bending, sitting, etc., within a valid range that the depth camera can collect.

We propose a novel method for estimating human body height from a single depth image. We first use a Kinect [32] to capture an RGB-D image, but discard its color data and only use the depth image to estimate the height, because we find the prediction result that only using depth image is better than using RGB or RGB-D image. Secondly, we segment the human torso from the depth image by using FCN [25]. In order to make the edge information more accurate, we enhance the depth image using high-frequency information. After obtaining the torso image, we propose a novel part-based intermediate representation. The torso image and depth image are input to the same network architecture to get the body parts segmentation that the human torso is divided into four local parts: head, upper body, thigh and calf. We verify that using this intermediate representation approach can significantly improve the accuracy of estimation. After that, we modify the fully connected layer of VGG16 [35] to make it more adaptable to our problem. Then the body parts segmentation and depth information are entered into this network to get intermediate estimation of lengths of the four body parts. After obtaining the intermediate representation above, we design a novel network architecture for the final height prediction.

Even though complex neural networks have very long training time, large number of parameters, and easy to overfit [6][15][20][30], generally, they can fit the data better than the shallow neural network and the prediction may be more accurate. So we propose an increasingly complex deep neural network which we called developing network architecture. At the beginning of the network, we only use a small number of convolutional layers to train, then output the predicted values through the fully connected layer. When the network basically converges, we add new convolutional layers, and continue to train the network. Repeating this process until the network can accurately estimate. Experiments show that the accuracy of network trained by this

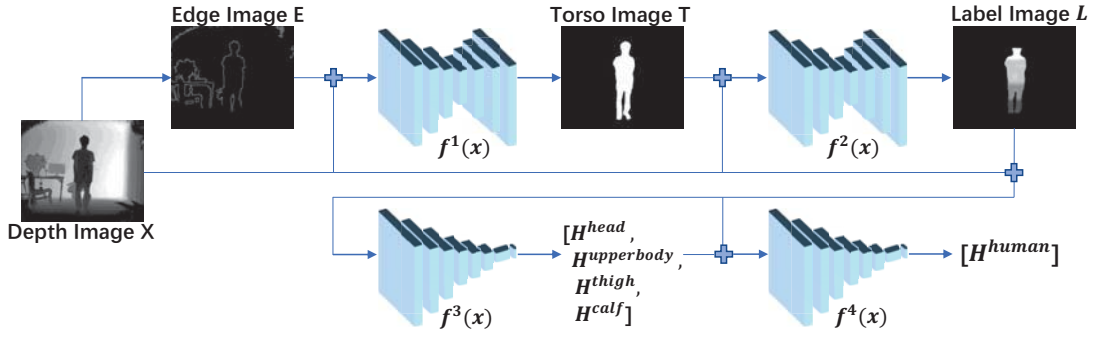


Figure 1. The workflow of our approach. We use a four-stage neural network to estimate human body height from a single depth image.

process is higher, compared with that of direct training. At the same time, for different characteristics of segmentation image and depth image, we adopt different pooling strategies and use the skip-connection structure [14] to transmit these information to each block without loss, see Figure 6, which further improve the accuracy of estimation significantly.

Our main working steps are as follows: Firstly, the human torso is segmented from the depth image, and then construct an intermediate expression, body parts segmentation image, which further divides the human torso into four local parts: head, upper body, thigh and calf. After that, the lengths of these four parts are predicted separately. Finally, a novel developing network architecture is devised and the network is trained by a hybrid pooling strategy. The above process is shown in Figure 1.

Our main contributions are as follows:

1. We construct a new dataset of human body heights including 2136 RGB-D images with ten postures such as standing, walking, sitting, bending, etc. The human body can be located at any position within the range that can be captured by a depth camera, Figure 2 shows some examples from our dataset.
2. To the best of our knowledge, this is the first method to predict human height from a single depth image. We only need a commodity depth camera without extra equipment, thus reduces the overall equipment expenditure for practical applications. By using only depth data, we achieve higher accuracy than using depth and RGB, see Section 4.4., which is an interesting fact contradictory to the case of 3D reconstruction where depth+RGB has been proved to be the better input [9] [11].
3. We verify how the intermediate representation can make the final network easier to learn. It proposes to construct the human body parts segmentation image and estimate the lengths of head, upper body, thigh, and calf respectively. Compared with the method without intermediate representation, the prediction accuracy is greatly improved.

4. We put forward a network architecture with a gradual complexity of iterations which can solve the difficult problem of network initialization and significantly improve the accuracy.

## 2. Related Work

In the field of human height estimation using images or videos, there are not too many previous work especially in recent years, and we have divided these methods into the following three groups.

### 2.1. Single-view Based Prediction

The single-view-based height prediction methods are relatively easier methods in current height measurement. Their methods make the prediction commonly based on camera calibration and reference substances, with an RGB or RGB-D image acquisition device to capture images from a certain angle. Penders et al. [28] propose a reference-based method of fixing the camera to a chosen position, keeping the distance between the subject who is required to stand close to the wall and the camera constant, thus getting the measured value of the distance between the head and the feet, finally converting the measured value into the actual distance with the reference measurement. Criminisi et al. [8] put forward to use the vanishing line and vanishing point method to calibrate the camera, thus eliminating the need for camera built-in parameters. Based on the [8], Lee [19] adds a cube in the image and use genetic algorithms to improve the robustness. In [2], [10] and [12], a heterogeneous method is presented that does not use any calibration or reference, but adopts a proportional relationship between body parts for estimation.

### 2.2. Multi-view Based Prediction

The multi-angle based prediction methods can be roughly categorized into multiple angles of shooting with a single camera, or taking multiple photos with a fixed camera, or shooting with multiple cameras at different angles. Three dimensional reconstruction is used to estimate human body data in [21] and [24]. They propose a three-dimensional

modeling through multi-angle photographs and then estimating the human body data through a cubic spline. Li et al. [23] use a home camera and a simple rotating disc to collect body images from different angles. Then perform a 3D reconstruction and refine the above model to estimate the human body data. Hung et al. [16] collect the front view, back view and side view of the human body, and then calculate the height of the human body by placing the standard reference.

### 2.3. Video-based prediction

Video-based method is also a common means of height prediction. Compared with the two methods above, this method can automatically and accurately segment the human body from the background, and then estimate the height. The collected video sequence be used to separate the background in [4], [17] and [18], then extract the feature points of the head and feet, and finally calculate the actual height by camera calibration or reference method. Li et al. [22] adopt a non-linear model to evaluate the focal distance, inclination and the height of camera, which removes the noisy interference during camera calibration. Shao et al. [33] use the moving objects in the tracking scene to recover the minimum calibration of the scene, and then adopt the single frame prediction method proposed by [8] to predict on each frame, and finally combine all the single-frame results together to obtain the final multi-frame prediction.

Although there have been many studies on estimating height from images or videos in recent years, the previous methods have some limitations. Most of the methods can only be used to measure postures such as walking and standing, or require the subject to stand at a specified position. Some methods need manual label of the head and feet, which is not fully automated and requires a lot of manpower. There are also others methods that use multiple photos or multiple devices, which cost more time and money. We are committed to research a fully automated height measurement method that requires only one depth camera for various posture, including extreme postures such as sitting and walking.

## 3. Method

This paper investigates how to estimate the height of a human body from a single depth image. In this section, we will show how to create data sets, how to establish intermediate expressions, and how to find an effective network architecture.

### 3.1. Data Set

Our first problem is to create a depth image dataset with height information. There are already some datasets containing human body information, such as W8-400 [27], RGB-D-T [29], etc., but in these data sets the human body is locat-

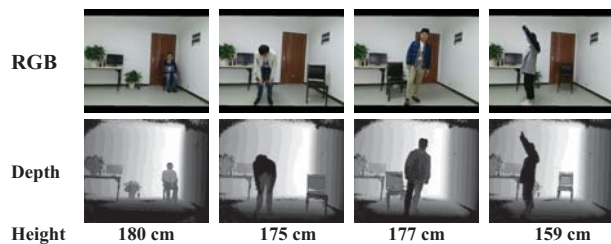


Figure 2. Some examples of our dataset. Our dataset consists of RGB images, depth images and the corresponding human height.

ed in the center of the image, or with just a simple posture such as upright, walking, or standing straight to the camera. However, these simple data cannot meet the needs of our real-life scenes, just like when we perform 3D reconstruction on the human body, the subject may be located anywhere in the image, and pose various postures, or during the medical height measurement, the patient may not be able to stand. The methods based on these data sets cannot be effectively applied. These problems are the foci and difficulties of the height prediction field, and also the motivation of our research.

We create a human body dataset with 2136 RGB-D images using a Kinect camera [32], but we only use the depth images, and the RGB images can be used by relevant research in the future. The data set consists of 10 postures, including walking, bending, sitting, etc., see Figure 8. There are 14 volunteers in our dataset. They can stand anywhere with arbitrary clothes. Their heights ranges from 158cm to 184cm, which covers a wide range of height [7]. Figure 2 shows some examples from our dataset.

Next, we need to consider how to organize the training data to ensure the network really establishes a connection between depth information and height, instead of connection between identity and height. To verify this, we extract a man and a woman from 14 people and put all of their 369 images into the test set, also called Strange-test, to avoid the network learning their identity information. For the other 12 people, each person randomly selects 5 images with a total of 60 into the test set, also called Familiar-test.

In summary, our dataset contains 2136 depth images equipped with their corresponding body height values, which is divided into a training set containing 1707 images and a test set containing 429 images.<sup>1</sup>

### 3.2. Intermediate Representation

As discussed below, we will consider how to establish an intermediate representation of height information from depth images, making it easier and more efficient for the network to estimate height information.

Segmentation of human body parts from depth images as intermediate representation has been well applied in pose

<sup>1</sup>Please visit our project pages at: <https://depth2height.github.io>

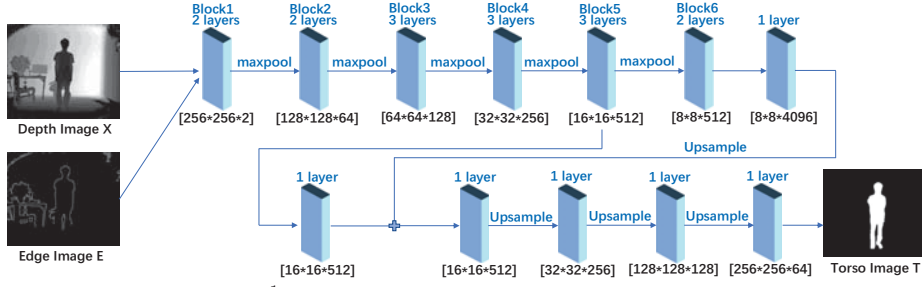


Figure 3. The network architecture of  $f^1(X)$ . The number below each block or layer represents their input size. Input a depth image  $X^D$  and an edge image  $E$  to output the corresponding torso image  $T$ .

estimation [26] [34]. Similarly, in this paper, we use the human body parts segmentation image within our intermediate representation.

In order to make the segmentation more suitable for our problems, we observe that the human head, upper body, thigh, calf are nearly rigid, and the height can be expressed as the sum of the four parts. More importantly, their relative positions can reflect the overall posture. Therefore, we segment the human torso into these four nearly rigid parts using depth images and torso images. In order to eliminate the interference of the hairstyle, we define the head part as between the eyebrow to the neck. We show some of our intermediate representation images in Figure 4.

Experiments show that using our intermediate representation can significantly increase the accuracy of height estimation, see Section 4.4. It provides three advantages:

1. We decompose the problem of human height estimation into estimating the height of four nearly rigid parts as the length of a rigid object is more predictable.
2. The topological structures and length proportion relationships between these four parts contain the posture information, which provides a powerful clue for height prediction.
3. Convolutional neural network (CNN) has a good performance in local perception [1]. Decomposing the problem into local small problems can take advantage of CNN and simplify complexity, which makes the final estimation more accurate and stable, see Table 1.

### 3.3. Network Architecture

In this section we will describe our network architecture for obtaining the intermediate representation.

In the first step, we need to segment the human torso image from the depth image. FCN [25], Unet++ [36], Mask r-cnn [13], and other methods [5] [31] have great performance in pixel-to-pixels image segmentation. We slightly modify the input so that we can accept the depth as input and output the human torso segmentation image. However, although these methods can generally segment the torso information, but cannot perform well at the edge of the human body, especially at the head and the feet area. We know

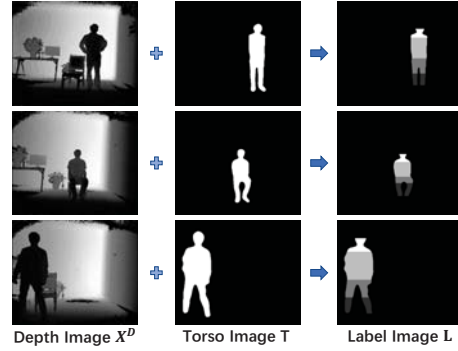


Figure 4. Some examples of our part-based intermediate representation. Input the depth image  $X^D$  and the torso image  $T$  into  $f^2(X)$  to get the label image  $L$  as our intermediate representation in the form of a segmentation of human torso into four parts.

that in the field of distance prediction, the determination of the starting point and the ending point is of paramount importance, see Figure 10. So we consider inputting the high frequency information of the depth image to the network to improve the prediction ability of the edge. Let the original image as  $X^D$  and the high frequency image as  $E$ . We use the canny operator [3] to extract the edge information:

$$E = \text{canny}(X^D) \quad (1)$$

Then we input the depth image  $X^D$  and the high frequency information  $E$  into the convolutional neural network  $f^1(X)$  as shown in Figure 3, and define the loss function as:

$$\mathcal{L} = \frac{1}{N} \sum_{i \in N} \|f_i^1(X^D, E) - T_i\|^2 \quad (2)$$

Among them,  $T$  is the human torso image,  $i$  every pixel and  $N$  is the total number of pixels in the torso image.

We adopt the same convolutional network architecture as the  $f^1(X)$  to design a new network  $f^2(X)$  that input depth image  $X^D$  and the torso image  $T$  to output the corresponding label image  $L$  as our intermediate representation. Similarly, we define the loss function of the network as:

$$\mathcal{L} = \frac{1}{N} \sum_{i \in N} \|f_i^2(X^D, T) - L_i\|^2 \quad (3)$$

$i$  every pixel and  $N$  is the total number of pixels in the label image  $L$ .



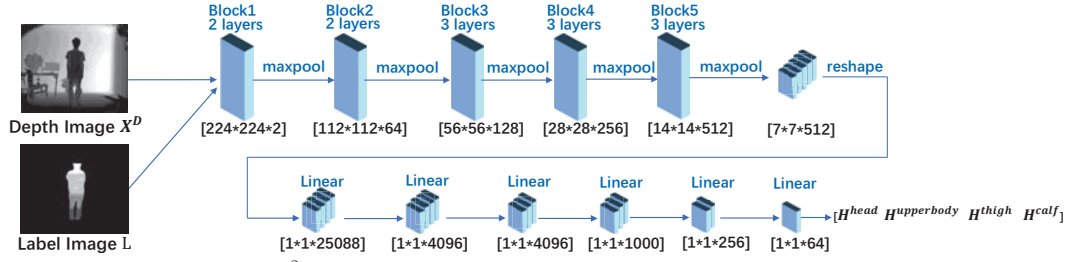


Figure 5. The network architecture of  $f^3(X)$ . We modify part of the full convolution layers and add three full connected layers based on VGG16 to input a depth image  $X^D$  and an edge image  $E$  then output the estimated lengths of these four parts.

After dividing the torso into four parts, we need to predict these four parts separately, and get their lengths  $H^{head}$ ,  $H^{upperbody}$ ,  $H^{thigh}$ ,  $H^{calf}$  of the head, upper body, thigh and calf in the image respectively. We design a new network architecture  $f^3(X)$  similar to VGG16 [4], but modify the fully convolution layer and add three full connected layers at the end of the network to make it more suitable for our problem. The network structure is shown in Figure 5. We enter the depth image  $X^D$ , and the label image  $L$  into the network, and get the estimated length of these 4 parts, namely:

$$[H^{head} H^{upperbody} H^{thigh} H^{calf}]^{1 \times 4} = f^3(X^D, L) \quad (4)$$

In this way, we construct our intermediate representation method of height prediction, a part-based segmentation image and the length of each part. In the next section, we will discuss how to use this intermediate representation for the final prediction and verify our proposed representation method is effective in Chapter 4.

### 3.4. Developing Network

In this section, we will discuss how to use our intermediate representation to estimate body height. We design a developing network architecture with a hybrid pooling approach to solve this problem. In large-scale and medium-scale convolutional blocks, it is still very important to obtain accurate depth information  $X^D$  and label information  $L$ , so we enhance a skip connection structure [14] to directly add these information to the output of the previous convolution-

al layer as input of the next convolutional layer. Initially, we simply let the depth information  $X^D$  and the label information  $L$  use the same pooling strategy. However, we find that even if the skip connection structure is added, the network accuracy does not increase significantly. Later, we observe that there are a lot of noise in the depth image  $X^D$ . It is easy to be interfered by these noises with maxpool, so we change the pooling mode to average pooling which can bring a certain sense of smoothness, and alleviate the interference from the local extreme pixels. We also find that for label image, however, maxpool is more suitable. So we propose a new network architecture based on VGG16 [35] with skip connection structure and a hybrid pool strategy. The network structure diagram  $f^4(X)$  is shown in Figure 6.

During the training process, the network is prone to overfitting. So we propose an increasingly complex network

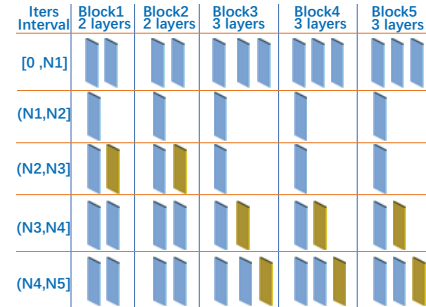


Figure 7. The change of our developing network structure with different iteration intervals. The blue layers are already working layers, and the yellow layers are newly added layers with the number of iterations.

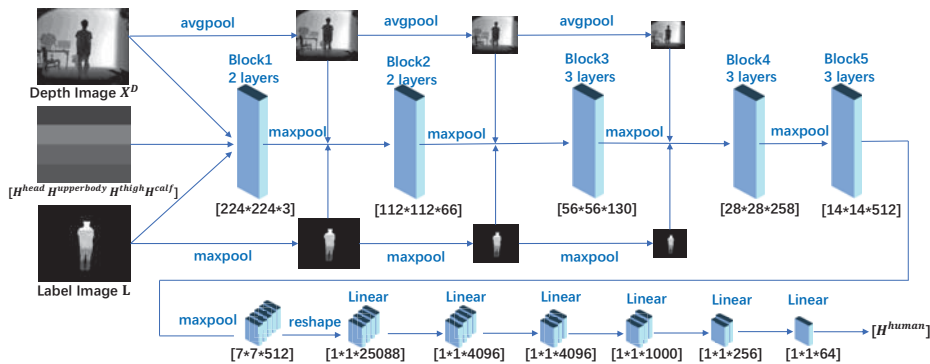


Figure 6. The network architecture of  $f^4(X)$ . We adopt different pooling strategies for the depth image  $X^D$  and the label image  $L$  to estimate human height.




Posture	Upright	Walking	Sitting	Bending	Arms raising Slightly	Unrolling Arms	Arms over Head	Waving Hands	Clapping	Having a Waistline
Input										
Truth	162 cm	162 cm	162 cm	162 cm	162 cm	162 cm	162 cm	162 cm	162 cm	162 cm
Estimate	162.11 cm	162.57 cm	161.50 cm	162.26 cm	161.77 cm	162.54 cm	161.91 cm	161.05 cm	161.49 cm	161.53 cm
Input										
Truth	175 cm	175 cm	175 cm	175 cm	175 cm	175 cm	175 cm	175 cm	175 cm	175 cm
Estimate	174.66 cm	174.71 cm	175.00 cm	175.08 cm	175.37 cm	175.07 cm	174.81 cm	174.61 cm	175.01 cm	175.52 cm
Input										
Truth	174 cm	158 cm	177 cm	175 cm	159 cm	161 cm	167 cm	164 cm	161 cm	184 cm
Estimate	173.88 cm	157.91 cm	176.94 cm	174.82 cm	158.93 cm	160.89 cm	166.92 cm	163.91 cm	160.92 cm	183.45 cm

Figure 8. Some estimation results. The first and second row are a female and a male volunteer from the Strange-test respectively. The third row shows volunteers from the Familiar-test. All subjects can be located anywhere and pose various postures.

In any case, our method can make accurate estimation only from a single depth image.

structure, i.e. developing network to avoid overfitting, as shown in Figure 7. We first pre-train with all convolutional layers until the number of iterations exceeds  $N_1$ . Then the model is saved, and only the first layer network of each block is reserved. When the iterations exceeds  $N_2$ , the second layer of the first and second block is added to training. Similarly, a new layer of the third, fourth and fifth blocks is added when the iterations exceeds  $N_3$  and  $N_4$ . It will continue to train until the iteration number reaches  $N_5$ .

We train these four subnetworks, i.e.,  $f^1(X)$ ,  $f^2(X)$ ,  $f^3(X)$  and  $f^4(X)$  one-at-a-time, see Figure 1. Our proposed network architecture can effectively use our intermediate representation to predict the height of human body, and has excellent performance in preventing over-fitting, see Figure 9 and Table 3.

## 4. Experiment

We conduct a series of experiments to validate our method, including whether to use intermediate representation, whether to adopt our developing network, and different input types. We also compare the accuracy with other methods. Experiments show that using intermediate representation and our network architecture has the highest accuracy over other methods. We show some results of our method in Figure 8.

### 4.1. Data Preparation and Parameter Settings

We resize 2136 depth images to  $256 \times 256$  and  $224 \times 224$ . In the networks  $f^1(X)$  and  $f^2(X)$ , we use images of size  $256 \times 256$ , in the network  $f^3(X)$  and  $f^4(X)$ , we use the size of  $224 \times 224$ . We train our network on the train set of 1707 images and test on the test set of 429 images.

The depths of the networks  $f^1(X)$  and  $f^2(X)$  are 21, and the other two are 19. The initial learning rate of the

networks  $f^1(X)$  and  $f^2(X)$  are set to 0.0001, and every 5 epochs we decrease the learning rate by a factor of 0.8. The other two are set to 0.0001, and for every 50 epochs will decrease the learning rate by a factor of 0.5. The batch size of network is set to 8.

We use the average relative error as evaluation index, which is defined as:

$$AverageRelativeError = \frac{1}{n} \sum_n |H_e - H_a| / H_a \quad (5)$$

$n$  is the number of samples in the test set.  $H_e$  is the estimated height, and  $H_a$  is the real height.

Correspondingly, we define the accuracy as:

$$Accuracy = 1 - AverageRelativeError \quad (6)$$

### 4.2. Validity of the Intermediate Representation

In this section, we will verify the validity of our intermediate representation proposed in Section 3.2. We conduct three sets of experiments: our method, the method without intermediate expression (denoted by M1), and the method with partial intermediate expressions (denoted by M2).

In our method, we completely follow the steps mentioned in Section 3.2 to get the intermediate representation. In M1, we only use  $f^1(X)$  to segment the image, and then input the obtained torso image  $T$  and depth image  $X^D$  into  $f^4(X)$  to estimate height. In M2, we input the label image  $L$  which is obtained by  $f^1(X)$  and  $f^2(X)$  sequentially, and the depth image  $X^D$  into  $f^4(X)$  to get the result. Table 1 shows the relative error of the three methods.

Method Name	Error
<b>Ours</b>	<b>0.90%</b>
M1	1.71%
M2	1.20%

Table 1. Average Relative Error of Our Method, M1, M2.

It can be seen from Table 1 that our error is 43% less than that of M1, and 19% less than that of M2. At the same time, we count the number of images of the three methods in different error intervals, as shown in Table 2.

Error Interval	Ours	M1	M2
$0 < \text{Error} \leq 1\%$	<b>273</b>	172	228
$1\% < \text{Error} \leq 2\%$	101	113	128
$2\% < \text{Error} \leq 3\%$	35	65	45
$3\% < \text{Error} \leq 4\%$	17	45	16
$4\% < \text{Error} \leq 5\%$	2	21	3
$5\% < \text{Error} \leq 6\%$	1	7	6
$6\% < \text{Error} \leq 7\%$	0	3	1
$7\% < \text{Error} \leq 8\%$	0	0	1
$8\% < \text{Error} \leq 9\%$	0	2	1
$9\% < \text{Error} \leq 10\%$	0	1	0

Table 2. Number of Samples within Different Error Intervals in Our method, M1 and M2.

It is clear that from Table 2 that our method has excellent performance in reducing the extreme value. 99.3% of all results have error lower than 4%, and 87.2% of them lower than 2%.

The result demonstrates that using our intermediate representation can significantly improve accuracy and reduce the extremely incorrect estimation.

#### 4.3. Effectiveness of the Network Architecture

In order to verify the validity of our network architecture proposed in Section 3.3, we conduct two groups of contrast tests (called M3, M4). In M3, rather than using the increasingly complex network architecture, we use all convolutional layers to train the network. In M4, we do not pre-train the network, but gradually restore the network architecture from the simplest architecture. Table 3 shows the relative errors of the three methods.

Method Name	Error
<b>Ours</b>	<b>0.90%</b>
M3	0.97%
M4	1.07%

Table 3. Average Relative Error of Our Method, M3 and M4.

As can be seen from Table 3, our architecture is better than M3 and M4 by 7.22% and 15.89% less error respectively.

We plot the curve of accuracy rate during the training process in Figure 9. In our implementation, we set  $N1=40000$ ,  $N2=60000$ ,  $N3=80000$ ,  $N4=100000$ ,  $N5=160000$ . Before the  $N1$  point, M3 adopts the same training method as ours. we observe that the two methods have the similar correct rates. After reaching  $N1$ , although the accuracy rate of our method decreases at the beginning, it will gradually increase with the recovery of the network structure, and the final accuracy even exceeds that of M3 and M4, which indicates that our architecture is effective.

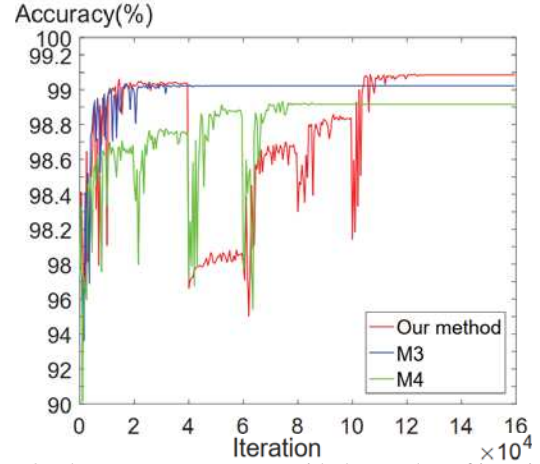


Figure 9. The accuracy rate curve with the number of iterations in four methods. In order to make the image clearer, we stretched the data within the range of [98,99.2] by 10 times.

#### 4.4. Comparison of Different Input Types

In this section we will show which input type is the best for our height estimation network. We conduct three sets of experiments using different inputs: RGB, RGB-D and depth only. For both RGB and RGB-D types, we only change the depth image in the input of network  $f^1(X)$ ,  $f^2(X)$ ,  $f^3(X)$  and  $f^4(X)$  to the corresponding image. Then we list the relative errors of these three input types in Table 4.

Input Type	Test Set	Strange-test	Familiar-test
RGB	1.35%	1.47%	0.64%
RGB-D	1.05%	1.13%	<b>0.56%</b>
<b>Depth(Ours)</b>	<b>0.90%</b>	<b>0.95%</b>	0.60%

Table 4. Average Relative Error of These Three Input Types on Test Set, Strange-test and Familiar-test.

Intuitively, RGB-D images contain more information than the other two types. This means using RGB-D images as input can output the most information. However, we find that solely using depth images generates the best result. We anticipate that the reason is because RGB data in the input will make the network establish a connection between identity and height, rather than estimating height from the image.

In order to verify our conclusion, we conduct experiments respectively on the two test sets: Strange-test and Familiar-test proposed in Section 3.1. We show the average relative error of each input type in these two test data sets in Table 4.

From Table 4 we find that the relative error of the input type using RGB image on the Strange-test is more than twice over the Familiar-test. This confirms our previous conclusion that using RGB images will enable the neural networks to learn a connection between identity and height.

Posture Name	Ours	Deák et al. [10]	CRIMINISI et al. [8]	Camera Calibration
Upright	<b>0.59%</b>	4.10%	1.63%	1.35%
Walking	<b>0.97%</b>	2.95%	7.41%	2.79%
Sitting	<b>1.00%</b>	3.58%	4.82%	25.50%
Bending	<b>2.19%</b>	9.92%	5.38%	28.53%
Arms raising Slightly	<b>0.78%</b>	4.38%	1.45%	1.20%
Unrolling Arms	<b>0.77%</b>	3.93%	1.52%	1.51%
Arms over Head	<b>0.91%</b>	4.86%	1.52%	1.48%
Waving Hands	<b>0.75%</b>	4.45%	1.54%	1.53%
Clapping	<b>0.69%</b>	5.50%	1.58%	2.91%
Having a Waistline	<b>0.73%</b>	4.39%	1.59%	2.97%
Total Average Error	<b>0.90%</b>	4.80%	6.44%	2.69%

Table 5. The Average Relative Error between Our method and Other Methods in Different Postures.

#### 4.5. Comparison to Other Methods

This section compares our method with other methods on our test set, including Deák et al. [10] based on the proportional relationship between body parts, CRIMINISI et al. [8] based on vanishing point and vanishing line, and the Kinect camera calibration method [32]. Table 5 shows the comparison results.

It can be seen from the comparison that the average error of our method on the whole test set is significantly better than other methods. Our average relative error is approximately 0.90 which can accurately extract the human body height from the image. Since the methods of CRIMINISI et al. [8] and Camera calibration can not cope well with the non-upright posture of the human body, in Table 5, we separately calculate the error of the four methods according to different postures of the human body. It is noticeable that ours is the best among the four methods in various postures.

At the same time, we analyze the error of other methods. Deák et al. [10] only relied on the ratio between the pixel length of the head and the pupillary distance, so its sensitivity to the gesture is small. But this method requires that a human face should be possibly perpendicular to the camera's optical axis, otherwise the ratio of the distance between the human eye and the length of the head will not be correctly extracted. Besides, due to the differences of individuals, it is not convincingly accurate to solve the height of the person through the statistical law. In the methods of CRIMINISI et al. [8] and Camera calibration, when a person walks, sits, or bends, the straight line distance between the top of the head and the bottom of the foot cannot be a good representation of the height of the human body.

Although these two methods have better performance in the upright pose compared to the aforementioned postures, they have the inherent problem in predicting body height from a single-view image. As shown in Figure 10, the distance  $h2$  between dash lines is the real height of the human body, and  $h1$  is the predicted height according to the image. Therefore, there is still about 1.5% error in the upright posture for both methods.

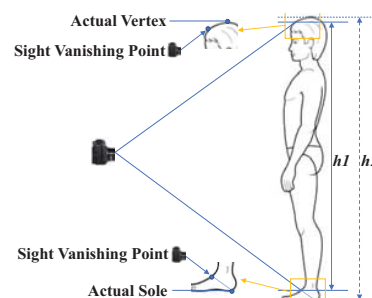


Figure 10. A failure case with estimating height only using a single-view image in [10] and [8].

#### 5. Conclusion and Future Works

We create a method for accurately and quickly estimating body height from a depth image based on increasingly complex network architecture. We propose an intermediate representation based on an effective body torso segmentation, which is automatically obtained by adding high-frequency information of the depth image into a FCN. We first predicts the lengths of each body parts respectively and eventually construct a developing and complex network for the final estimation which effectively suppresses the model over-fitting phenomenon. Our method can cope with sitting, bending, walking and many other postures, and the accuracy rate can even reach at 99.1%.

In future we may enrich our current dataset with more subjects and refine the human body segmentation based on semantic bioinformation which we believe will further improve the accuracy. For future direction, we would like to explore more non-contact measuring techniques of geometric and physical units such as weight and density using optimized deep learning with various inputs.

#### 6. Acknowledgements

This work was supported by the grant of Science Foundation of Hunan Province(No.2018JJ3064), National Science Foundation of China(No.61303147). We gratefully acknowledge NVIDIA for GPU donation. We thank Dan Yin, Wei Cai and Zeyu Liu for their help on dataset preparation.



## References

- [1] Angelos Amanatiadis, Vasileios G Kaburlasos, and Elias B Kosmatopoulos. Understanding deep convolutional networks through gestalt theory. In *2018 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE, 2018. 4
- [2] Chiraz BenAbdelkader and Yaser Yacoob. Statistical body height estimation from a single image. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–7. IEEE, 2008. 2
- [3] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 4
- [4] Yu Chai and Xiaojing Cao. A real-time human height measurement algorithm based on monocular vision. In *2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pages 293–297. IEEE, 2018. 3
- [5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4
- [6] Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015. 1
- [7] NCD Risk Factor Collaboration et al. A century of trends in adult human height. *Elife*, 5:e13410, 2016. 3
- [8] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision*, 40(2):123–148, 2000. 2, 3, 8
- [9] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(3):24, 2017. 2
- [10] A Deák, O Kainz, M Michalko, and F Jakab. Estimation of human body height from uncalibrated image. In *2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 1–4. IEEE, 2017. 2, 8
- [11] Yanping Fu, Qingan Yan, Long Yang, Jie Liao, and Chunxia Xiao. Texture mapping for 3d reconstruction with rgb-d sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4645–4653, 2018. 2
- [12] Ye-Peng Guan. Unsupervised human height estimation from a single image. *Journal of Biomedical Science and Engineering*, 2(06):425, 2009. 2
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5
- [15] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1389–1397, 2017. 1
- [16] Patrick Chi-Yuen Hung, Channa P Witana, and Ravindra S Goonetilleke. Anthropometric measurements from photographic images. *Computing Systems*, 29:764–769, 2004. 3
- [17] Erno Jeges, Istvan Kispal, and Zoltan Hornak. Measuring human height using calibrated cameras. In *2008 Conference on Human System Interactions*, pages 755–760. IEEE, 2008. 3
- [18] István Kispál and Ern Jeges. Human height estimation using a calibrated camera. In *Proc. CVPR*, 2008. 3
- [19] Kual-Zheng Lee. A simple calibration approach to single view height estimation. In *2012 Ninth Conference on Computer and Robot Vision*, pages 161–166. IEEE, 2012. 2
- [20] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016. 1
- [21] Jie Li, Mingui Sun, Hsin-Chen Chen, Zhaoxin Li, and Wenyan Jia. Anthropometric measurements from multi-view images. In *2012 38th Annual Northeast Bioengineering Conference (NEBEC)*, pages 426–427. IEEE, 2012. 2
- [22] Shengzhe Li, Van Huan Nguyen, Mingjie Ma, Cheng-Bin Jin, Trung Dung Do, and Hakil Kim. A simplified nonlinear regression method for human height estimation in video surveillance. *EURASIP Journal on Image and Video Processing*, 2015(1):32, 2015. 3
- [23] Zhaoxin Li, Wenyan Jia, Zhi-Hong Mao, Jie Li, Hsin-Chen Chen, Wangmeng Zuo, Kuanquan Wang, and Mingui Sun. Anthropometric body measurements based on multi-view stereo image reconstruction. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 366–369. IEEE, 2013. 3
- [24] Yingying Liu, Arcot Sowmya, and Heba Khamis. Single camera multi-view anthropometric measurement of human height and mid-upper arm circumference using linear regression. *PloS one*, 13(4):e0195600, 2018. 2
- [25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 4
- [26] Natalia Neverova, Christian Wolf, Florian Nebout, and Graham W Taylor. Hand pose estimation through weakly-supervised learning of a rich intermediate representation. *Pre-print: arxiv*, 151106728, 2015. 4
- [27] Tam V Nguyen, Jiashi Feng, and Shuicheng Yan. Seeing human weight from a single rgb-d image. *Journal of Computer Science and Technology*, 29(5):777–784, 2014. 3
- [28] Bas Penders, Ralph Brecheisen, Angèle Gerver, Geertjan van Zonneveld, and Willem-Jan Gerver. Validating paediatric morphometrics: body proportion measurement using photogrammetric anthropometry. *Journal of pediatric endocrinology and metabolism*, 28(11-12):1357–1362, 2015. 2
- [29] Christian Pfitzner, Stefan May, and Andreas Nüchter. Body weight estimation for dose-finding and health monitoring of

- lying, standing and walking patients based on rgb-d data. *Sensors*, 18(5):1311, 2018. 3
- [30] Aaditya Prakash, James Storer, Dinei Florencio, and Cha Zhang. Repr: Improved training of convolutional filters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10666–10675, 2019. 1
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4
- [32] Microsoft Kinect sensors for Windows SDK. Available online: <https://docs.microsoft.com/en-us/previous-versions/windows/kinect>. 1, 3, 8
- [33] Jie Shao, Shaohua Kevin Zhou, and Rama Chellappa. Robust height estimation of moving objects from uncalibrated videos. *IEEE Transactions on Image Processing*, 19(8):2221–2232, 2010. 3
- [34] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304. Ieee, 2011. 4
- [35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 5
- [36] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018. 4