On the Feasibility of 3D Model-Based Forensic Height and Weight Estimation

Neerja Thakkar and Hany Farid
University of California, Berkeley
Berkeley, CA USA
{nthakkar,hfarid}@berkeley.edu

Abstract

Forensic DNA analysis has been critical in prosecuting crimes and overturning wrongful convictions. At the same time, other physical and digital forensic identification techniques—used to link a suspect to a crime scene—are plagued with problems of accuracy, reliability, and reproducibility. Flawed forensic science can have devastating consequences – the National Registry of Exonerations identified that flawed forensic techniques contribute to almost a quarter of wrongful convictions in the United States. Even some of the most basic, general-purpose forensic techniques for measuring a person's height and weight are unreliable. We propose using recent advances in 3D body-pose estimation to estimate height and weight from a single, unconstrained image. The reliability of this method is assessed using large-scale simulations and an in-the-wild dataset, bounding the expected accuracy with which height and weight can be estimated, and providing a road map for further improvements.

1. Introduction

In 2005, the U.S. Congress authorized the National Academy of Sciences (NAS) to conduct a study of forensic science. Published in 2009, the committee's report [20] called for a broad and deep restructuring of how forensic techniques are validated and applied, and how forensic analysts are trained and accredited. One of the report's key findings was "[w]ith the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source." A decade later, Judge Harry Edwards, co-chair of the original committee, wrote "We are still struggling with the inability of courts to assess the efficacy of forensic evidence. When a forensic expert testifies about a method that has not been found to be valid and reliable, the expert does not know what he does not know and cannot explain the limits of the evidence. This is unacceptable" [12].

Figure 1. Can we accurately determine how tall this person is?

With digital devices in nearly every hand, photographic evidence is playing an increasingly important role in identifying people. The field of photographic forensic identification, however, is under-studied and, as the NAS found, riddled with flawed forensic techniques (e.g., [22]).

In the summer of 2008, for example, George Powell III was identified as a possible suspect in a string of convenience store armed robberies, Figure 1. A store clerk initially told police the robber was approximately 167 cm (5 ft, 7 in), and went on to identify Powell in a lineup. Powell stands at 190 cm (6 ft, 3 in). A former police officer, turned expert witness, testified in court that a photogrammetric analysis of video surveillance showed the robber was at least 185 cm tall. Powell was convicted and sentenced to 28 years in prison.

After his conviction, Powell's family hired two other experts, one of whom concluded the video surveillance showed the robber was 171 cm, and the other bracketed the robber's height at between 167–176 cm. In light of these measurements, the original expert revisited his findings and adjusted his estimate to a range of 178–185 cm. Due in part

to these photogrammetric inconsistencies, Powell's conviction was vacated in 2018, and he was granted a new trial.

The field of photogrammetry, dating back to 1867, is mature and the underlying techniques are well understood [18, 26]. Yet, as the Powell case showed, even the seemingly straight-forward task of measuring a person's height can be wildly inaccurate. One of the most significant challenges in assessing even the most basic measurement of a person's height is contending with the inherent loss of information resulting from a 3D to 2D image projection, along with often low image resolution, perspective distortion, and the reality of measuring a person's height when they are not necessarily standing up straight.

Many factors add to the complexity of measuring the human body. Due to spinal compression throughout the day, a person's reference height can change by as much as 1.9 cm [13]. A person's apparent height can change by as much as 6 cm while they are walking [9]. This type of apparent height difference is even more exaggerated for extreme poses or actions. Weight similarly fluctuates in the span of weeks as well as throughout a single day. And, clothing, shoes, and head-wear can further complicate both height and weight measurements.

While not all of these complexities can be easily overcome, we hypothesize that recent advances in 3D human pose estimation [15, 24] can reduce some of the ambiguities and uncertainty inherent to height and weight estimation from a single, reference-free image. With this approach, and through large-scale simulations and in-the-wild experiments, we evaluate the accuracy and reliability with which a person's height and weight can be measured. We also evaluate the accuracy and reliability as a function of body size and pose, camera angle, and image resolution and quality.

2. Background

The heights of adult women/men in the U.S. are normally distributed with a mean of 161/175 cm, and a standard deviation of 7.0/7.4 cm [1]. If we were to simply estimate a person's height as this average, gender-specific height, the average height estimation error for women/men would be 5.6/5.6 cm, or 3.5%/3.2%.

Similarly, the average U.S. adult female/male weight is 78.7/90.8 kg with a standard deviation of 19.7/19.8 kg [1]. If we were to estimate a person's weight as this average, gender-specific weight, the average weight estimation error for women/men would be 15.8/15.8 kg, or 20.0%/17.3%. We report these numbers as a baseline against which height and weight estimation should be considered.

Many classic approaches to forensic height estimation require the presence of a reference object in the scene (e.g., a door frame of known size). Criminisi et. al [9, 8], for example, leverage insights from projective geometry and computer vision to extract height measurements from images and videos in which the person in question is standing next to a reference object of known height. The authors note that when a person is standing upright they can measure a person's height, although no large-scale studies were performed to determine the accuracy or reliability. In contrast, our proposed 3D body pose estimation should be able to accommodate different body poses, a necessary, but not necessarily sufficient step to accurate height estimation.

A few approaches for estimating human height that do not rely on reference objects have been proposed. BenAbdelkader et. al [4] combine classical single-view metrology with statistical knowledge of human anatomy to estimate height. This approach is significantly less accurate than reference-based, single-view metrology, and is only slightly better than guessing a gender-specific average height. Zhu et. al [27] use priors learned by a neural network to perform geometric camera calibration and recover the absolute scale of a scene from a single image. With a mean absolute error (MAE) of 8.3/12.1 cm (for people in neutral/non-neutral poses), this approach is worse than guessing a gender-specific average height. Bieler et. al [5] use explicit knowledge of gravity to measure a person's height from a video sequence. With a MAE of 3.9 cm, this approach is better than, for example, guessing a gender-specific average height. This approach, however, is limited to video in which the person being measured is in free fall, subject to the gravitational force.

There have also been a few approaches to estimating weight from images and video. Velardo and Dugelay [25] estimate weight from seven, manually-extracted, anthropomorphic measurements. Using a pair of frontal and side view images, with a known reference object, they achieve a MAE of 7.2 kg. Arigbabu et. al [3] estimate weight by training a feedforward neural network with 13 measurements of the human body across multiple video frames, yielding a MAE of 6.4 kg. Nguyen et. al [21] estimate weight from a single RGB-D image, incorporating color, depth, and gender information in their estimate, yielding a MAE of 4.6 kg. In contrast to these approaches, our method uses a single image in the absence of a reference object.

3. Methods

We begin by describing our forensic measurement pipeline. The process starts with estimating a 3D model from a single image [24], from which direct height and weight measurements are made. Although the resulting 3D model is estimated in real-world units, we describe three alternate approaches for determining absolute scale. The first approach can only be used in simulation, the second approach can be used in simulation and in real-world forensic scenarios with a known suspect, and the third approach can be used in all real-world scenarios. We then describe the creation of a large-scale simulated dataset that allows us

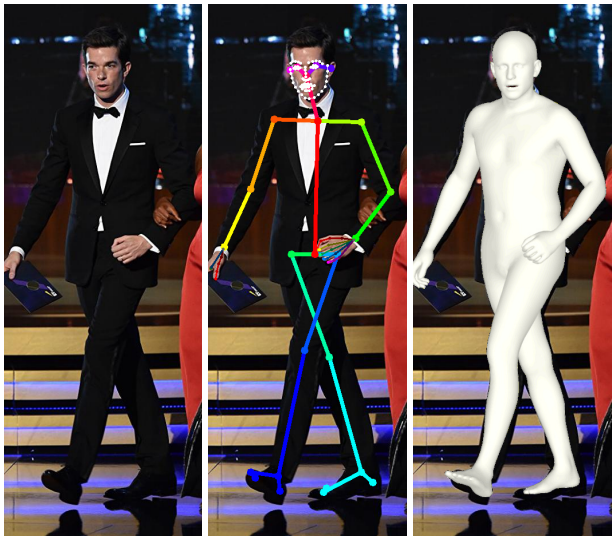

Figure 2. Overview of 3D model fitting on a sample image from the *IMDB-23K* dataset. Shown from left to right is: the input image; the estimated 2D skeletal keypoints; and the estimated 3D model. The 3D model is then reposed into a neutral pose, from which height is estimated as the distance from the top of the head to the plane of the feet, and a proxy for weight is estimated as the volume of the 3D model.

to evaluate the accuracy and reliability of measuring height and weight across a broad range of imaging and human-body configurations. We also apply our height estimation method to images from an in-the-wild dataset, allowing us to evaluate accuracy in more challenging real-world forensic scenarios.

3.1. 3D pose estimation

The SMPLify-X optimization approach is used to fit 3D SMPL-X models to a single image, capturing 3D pose, body shape, and expression [24]. The SMPL-X model, $M(\theta, \beta, \psi)$, is defined by three sets of parameters: the body pose, θ, the body, face, and hand shape β, and the expression ψ. As part of the shape information, the β parameter contains metric reconstruction information in real-world units. The SMPL-X model extends the widely-used SMPL model [17], allowing for articulated hands and more accurate and detailed face modeling.

The first step in fitting a 3D SMPL-X model to an image consists of automatically detecting 2D body, face, hand, and feet keypoints, Figure 2, using OpenPose [6]. The full 3D body model is then estimated by optimizing for the parameters θ, β, and ψ by minimizing the difference between the 2D keypoints and the posed 3D model keypoints reprojected into 2D [24]. This optimization incorporates several priors on human-body shape and pose. The body pose prior, VPoser, is represented as a 32D latent vector [24]. VPoser can also be used to interpolate between poses or generate novel valid human poses. We, for example, use VPoser in our simulations to generate a wide range of body poses, Figure 3.

In order to estimate height, the estimated 3D model is reposed into a neutral pose, Figure 3 (top right) from which height is measured as the distance from the top of the head to the plane formed by three points on the bottom of the feet. The volume of the 3D model, computed using Python's trimesh package, is used as a proxy for weight.

3.2. Scale estimation

Although the estimated body model is estimated in real-world units, we will see that this metric reconstruction can be highly inaccurate, even though the underlying pose is quite accurate. As such, we describe three different approaches to extracting more accurate metric measurements.

Working on the assumption that a scene may not always contain reference objects of known size, we leverage the fact that the adult inter-pupillary distance (IPD) is relatively similar for women and men [11]: the average adult IPD for women is 6.17 cm with a standard deviation of 0.36 cm, and 6.40 cm for men with a standard deviation of 0.34 cm.

Clinically, IPD is measured as the distance between the center of the two pupils as a participant is looking directly forward. Because our 3D models do not have pupils, we measure the center of the eye as the midway point between the left and right corners of the eye. We observe empirically that this definition of IPD is slightly larger than the clinical definition, and therefore scale all measured IPDs by 0.975.

While using an average IPD will lead to added uncertainty in our measurements, when a surveillance photo is being compared to a specific suspect, then the suspect's measured IPD can be used instead. This scenario is mimicked in our simulations by measuring the IPD of both the ground truth 3D model and the fitted 3D model, from which the fitted model can be appropriately scaled.

In order to disentangle the impact of the underlying 3D model estimation and the resolution of the scale ambiguity, we adopt, in our simulations, another approach to estimating absolute scale. In particular, we align the estimated 3D model to the ground-truth 3D model using coherent point drift (CPD) [19]. CPD is a point-set registration algorithm that estimates the 3D rotation, isotropic scale, and translation between two arbitrary point clouds (in our case, the point clouds correspond to the vertices of the underlying 3D models). After aligning the estimated 3D model to the ground-truth model, the height and weight can be estimated in the units of the ground-truth model.

3.3. Datasets

We describe the creation of a large-scale simulated dataset, allowing us to assess the reliability of height and weight estimation across a range of body sizes and poses,

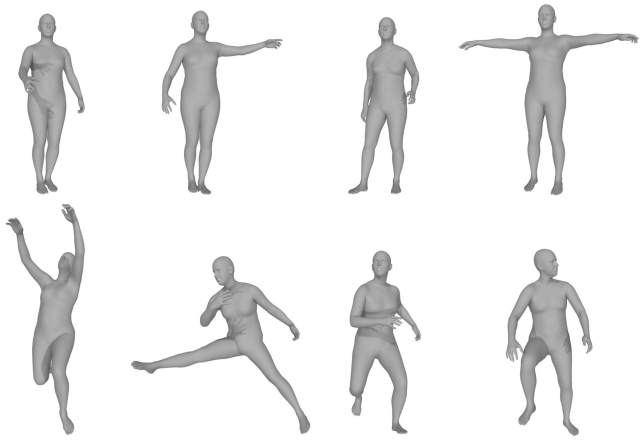

Figure 3. Representative examples of neutral (top) and action (bottom) poses used in our simulated dataset.

camera angles, and image resolutions and qualities. We also analyze images from an in-the-wild dataset allowing us to assess the reliability in real-world imaging situations where we must rely on only the average IPD to resolve the scale ambiguity.

3.3.1 Simulation

We created a large-scale, simulated dataset with known ground-truth height and weight (i.e., volume). This dataset is constructed from 3D SMPLX models of varying shape and pose, rendered with a range of camera angles, and post-processed with a range of resolutions and compression qualities. By starting with the same underlying 3D model used by the SMPLify-X pose estimation, we can disentangle the errors introduced by the 3D pose estimation and the scale disambiguation.

More specifically, our dataset is constructed by sampling three body shapes (small, medium, and large) from the SMPLX shape space. VPoser [24] is used to position the 3D model into one of 12 random poses, a subset of which is shown in Figure 3. For purposes of later analysis, these poses are categorized into neutral or action. These $3 \times 12 = 36$ posed models are rendered with a virtual camera positioned at one of 11 azimuths ($-150°$ to $150°$ in steps of $30°$) and one of 4 elevations ($0°$ to $54°$ in steps of $18°$), for a total of $1,584$ rendered images at a resolution of 800×800 pixels.

In order to assess the impact of image resolution and quality, each rendered image, at an azimuth and elevation of $0°$, is downsized to 400×400 and 200×200 pixels, and compressed with one of six JPEG qualities (100% (highest) to 0% (lowest) in steps of 20 using the Python Imaging Library, PIL). Each downsized and compressed image was then processed in the same way as the full resolution, uncompressed image.

3.3.2 In-the-wild

We use a subset of the *IMDB-23K* dataset [14] to evaluate the accuracy of height estimation in challenging real-world images. This dataset consists of images of celebrities annotated with their ground-truth height (but not weight). As with our simulated dataset described above, these images are categorized into two broad categories: neutral images consist of those where the person is standing facing the camera and with their entire body visible; and sitting/action images consist of non-neutral poses or images where a portion of the body is occluded. We selected a random subset of the *IMDB-23K* test dataset, consisting of 869 images in total, broken down as follows: 449 neutral female, 208 neutral male, 120 action/sitting female, and 92 action/sitting male.

4. Results

We begin by describing the accuracy with which height and weight can be estimated in simulation (using the dataset described in Section 3.3.1) for neutral and action poses, for a range of camera orientations, and for both the baseline SIMPLify, coherent point drift (CPD), and inter-pupillary distance (IPD) absolute scale estimation. The sensitivity to image resolution and compression is then explored. We then describe the height estimation accuracy for the in-the-wild dataset (Section 3.3.2). Finally, we briefly explore combining a classic single-view metrology analysis of height with the proposed 3D human modeling.

4.1. Simulation

4.1.1 Height

Shown in Figure 4(a) is the accuracy with which height can be estimated for neutral poses using only the SIMPLify metric information. These errors are reported as absolute percent deviation from the ground-truth height. The table rows correspond to the camera azimuth and the columns correspond to the camera elevation. Each cell is color coded proportional to the magnitude of the error, with higher saturation values corresponding to larger errors (all tables in Figure 4 use the same color-coding scheme on the same absolute scale). Shown in Figure 4(b) is the accuracy for action poses.

By comparison to an average error of 18.3% across neutral and action poses and across all camera orientations, if we were to guess an average, gender-specific height, the average height estimation error would be 3.5% (i.e., 6 cm for a 170 cm tall person). This shows that the metric information from SIMPLify is not sufficient to accurately measure metric height.

SMPLify (baseline)

(a) neutral poses

	0	18	36	54
[-150,-120]	5.6	7.3	16.7	30.9
[-90,-60]	5.9	6.1	18.0	39.9
-30	12.8	11.3	19.9	45.1
0	12.4	12.1	17.2	38.1
30	14.8	10.2	14.3	33.7
[60,90]	9.7	7.3	11.6	25.6
[120,150]	6.9	4.6	9.7	18.0

(b) action poses

0	18	36	54
11.1	6.0	13.3	23.3
11.6	14.9	26.6	52.4
15.7	18.8	29.5	66.0
13.5	18.4	30.2	67.4
11.5	16.1	28.8	67.8
8.8	11.6	23.8	42.1
7.7	4.1	9.7	25.4

CPD scale disambiguation

(c) neutral poses

	0	18	36	54
[-150,-120]	2.2	1.7	2.6	4.5
[-90,-60]	0.9	1.0	1.9	6.7
-30	0.5	0.8	2.0	6.5
0	0.5	0.8	2.2	4.8
30	0.6	0.5	1.8	4.3
[60,90]	0.9	0.8	1.2	5.2
[120,150]	1.5	1.1	2.1	7.6

(d) action poses

0	18	36	54
2.5	1.9	2.0	8.2
1.3	1.7	3.7	7.9
1.2	1.9	2.6	4.1
1.0	1.5	2.8	6.2
1.0	1.3	2.7	6.7
0.9	1.3	2.8	4.4
1.3	1.0	1.6	5.9

IPD scale disambiguation

(e) neutral poses

	0	18	36	54
[-150,-120]	4.3	2.8	5.0	6.7
[-90,-60]	2.7	3.7	6.7	14.0
-30	2.3	2.8	5.3	16.3
0	2.7	2.2	5.0	12.8
30	3.1	2.0	4.2	9.8
[60,90]	3.2	2.9	5.3	11.5
[120,150]	3.9	3.1	4.8	6.2

(f) action poses

0	18	36	54
7.4	3.9	3.6	7.5
4.8	4.7	7.8	16.0
3.4	5.2	9.8	22.9
3.4	5.4	11.0	26.0
3.1	4.5	12.0	25.2
2.8	3.5	9.5	18.1
4.3	3.2	4.5	7.8

Figure 4. Shown in each table is the height estimation error (%) as a function of the camera azimuth/elevation (rows/columns) for baseline SMPLIify, CPD-, and IPD-based scale disambiguation, and for neutral and action poses. Each cell is color coded proportional to the magnitude of the error, with higher saturation values corresponding to larger errors. In a real-world scenario, guessing a gender-specific, average height would yield an average height error of approximately 3.4%.

Shown in Figure 4(c) is the accuracy with which height can be estimated for neutral poses using the full-reference coherent point drift (CPD) scale disambiguation. With an average error of only 0.5%, height estimation is most accurate at a camera azimuth and elevation of $0°$. With this camera configuration and at a height of 170 cm (5 ft, 7 in), for example, the height of a person in a neutral pose can be estimated to within 0.85 cm (0.33 in), considerably more

(a) CPD

	0	18	36	54
[-150,-120]	28.0	25.9	28.8	40.6
[-90,-60]	23.3	27.5	32.0	35.6
-30	19.1	20.4	24.6	28.2
0	20.6	16.4	21.1	25.7
30	20.6	16.7	18.8	23.0
[60,90]	20.0	23.5	27.8	38.4
[120,150]	25.6	26.3	29.5	47.3

(b) IPD

0	18	36	54
17.4	20.0	21.8	33.8
19.5	21.1	20.9	35.1
16.4	16.2	15.7	41.1
15.4	13.6	15.8	49.6
15.1	15.2	14.9	46.8
17.7	18.9	16.8	26.7
19.4	20.2	19.5	29.1

Figure 5. Shown in each table is the volume (i.e., weight) estimation error (%) as a function of the camera azimuth (rows) and elevation (columns) for CPD- and IPD-based scale disambiguation, and averaged over all neutral and action poses. In a real-world scenario, guessing a gender-specific, average weight would yield an average weight error of approximately 18.7%.

accurate than SIMPLify's baseline or guessing a gender-specific average height. As the camera azimuth and elevation move away from a direct view, accuracy declines, albeit fairly gracefully, eventually breaking down when the camera elevation exceeds $54°$. This is because the more extreme camera angles make it increasingly difficult to accurately and completely extract all 2D skeletal features used in the SMPLify-X 3D model fitting, Section 3.1.

Shown in Figure 4(d) is the accuracy with which height can be estimated for action poses, again, using the full-reference CPD scale disambiguation. At relatively small camera azimuths between $[-30, 30]$ and elevations $[0, 18]$, the average accuracy is 1.3%, as compared to 0.6% for neutral poses. This increase in error is due to the difficulty in accurately extracting 2D skeletal features in action poses. As with neutral poses, overall accuracy continues to degrade as the camera azimuth and elevation increases.

Shown in Figure 4(e) is the accuracy with which height can be estimated for neutral poses, using, this time, the measured inter-pupillary distance (IPD) to resolve the scale ambiguity. As expected, there is an overall increase in errors with this less precise scale estimation. Across all camera azimuths and elevations, the IPD-based error is 5.5%, as compared to an average CPD-based error of 2.5%. At relatively small camera azimuths between $[-30, 30]$ and elevations $[0, 18]$, however, the average error is only 2.5%. As before, we see the same breakdown in accuracy for more extreme camera angles, and action as compared to neutral poses, Figure 4(f).

4.1.2 Volume/Weight

Shown in Figure 5(a) is the accuracy with which volume can be estimated from neutral and action poses using the full-reference CPD scale disambiguation. These errors are

reported as the absolute percent deviation from the ground-truth volume. As in Figure 4, the table rows and columns correspond to the camera azimuth and elevation. Each cell is again color coded proportional to the magnitude of the error.

Even with a direct camera view (azimuth and elevation of $0°$), the average error in estimating volume is 20.6%. By comparison, if we were to guess an average, gender-specific weight, the average weight estimation error would be on the order of 19%, (Section 2). As with height, these errors increase with increasing camera azimuth and elevation. Using an IPD-based scale disambiguation, yields similarly poor volume estimation.

The failure to accurately estimate volume is due to the fact that the underlying SMPLify-X model estimation only uses the extracted 2D skeletal keypoints to estimate pose, with no consideration to the underlying pixel-level appearance. As a result, the pose estimation is biased towards its prior of average-sized people, and appears unable to accurately capture overall shape. This limitation does not, however, significantly impact the ability to measure height.

Because we are unable to accurately estimate volume/weight in these simulations, we abandon trying to estimate weight from an in-the-wild dataset. We do, however, discuss in Section 5 possible remediations that may be considered to improve weight estimation.

4.1.3 Resolution and Compression

With an image resolution of 800×800, 400×400, and 200×200, and with the most direct camera view (camera azimuth and elevation of $0°$), the average error using CPD-based scale disambiguation is 0.7%, 0.7%, and 0.4%. The average error using IPD-based scale disambiguation over these three resolutions is 3.0%, 3.1%, and 3.2%. Over changes in resolution spanning a factor of four, there is little impact of image resolution on overall accuracy.

With a JPEG compression of 100% (best quality) to 20%, the average error using CPD-based scale disambiguation remains low, with errors between 0.56% and 0.67%. Only at the lowest quality of 0% is there a significant impact of compression, with the error increasing to 2.1%.

JPEG compression had a slightly larger impact when using IPD-based scale disambiguation. With a compression quality between 100% to 40%, the average errors are between 2.9% and 3.1%. At lower compression qualities of 20% and 0%, the errors rose to 3.4% and 4.6%.

JPEG compression has a more significant impact than resolution because the compression artifacts interfere with the ability to extract reliable 2D skeletal keypoints.

4.2. In-the-wild

Shown in Figure 6 is the accuracy with which height can be estimated from 869 *IMDB-23K* images, Section 3.3.2, (this dataset does not contain ground-truth weight). In these real-world images, only the average IPD-based scale disambiguation is applicable. These results are separated (top to bottom) based on gender. Shown in each panel is the absolute prediction error (cm) as a function of the ground-truth height, binned by 2.5 cm. The values specified in each bar correspond to the number of images at each ground-truth height.

For neutral poses, the minimal error is 0.003 cm, but increases to as much as 25.2 cm. Across all female/male heights, the average neutral error is $6.4/7.1$ cm, or $3.8/3.9\%$, as compared to the average IPD-based simulation errors of 2.7% for neutral poses. The average error across action and sitting poses for female/male heights is $7.6/8.2$ cm, or $4.6/4.6\%$ compared to an average IPD-based simulation errors of 3.4% for action poses.

We note that the errors systematically increase as heights increase away from the average female/male height of $161/175$ cm. We posit this is because the inter-pupillary distance (IPD) scales with height. To test this hypothesis, we linearly scaled the reference IPD proportional to height. While this is unrealistic in real-world forensic analysis, we see in Figure 7 that the resulting errors are significantly reduced to an average error of 3.9 cm for neutral poses and 5.1 cm for action and sitting poses. By comparison, guessing the average, gender-specific height would yield an error of 5.6 cm.

With the appropriate IPD, height estimation is reasonable for neutral poses, but is only slightly better than guessing for more complex poses. Combined with the results from our large-scale simulations, Section 4.1, we see the rate limiting step to accurate height estimation, for neutral poses and direct camera orientations, is not the 3D pose estimation, but rather the scale disambiguation. For more complex poses and off-axis camera orientations, both the 3D pose estimation and the scale disambiguation introduce significant errors.

4.3. Augmented Single-view Metrology

We have shown that modeling 3D pose can improve height estimation when a person is not standing perfectly upright and directly facing the camera. Resolving the scale ambiguity from this 3D model with the average human inter-pupillary distance (IPD), however, results in significant errors in real-world scenarios. On the other hand, single-view metrology is better at resolving the scale ambiguity because of the assumption of a known reference object, but has not previously incorporated 3D pose estimation and correction. Here, we describe an example of augmenting a classic, single-view metrology estimate of height with

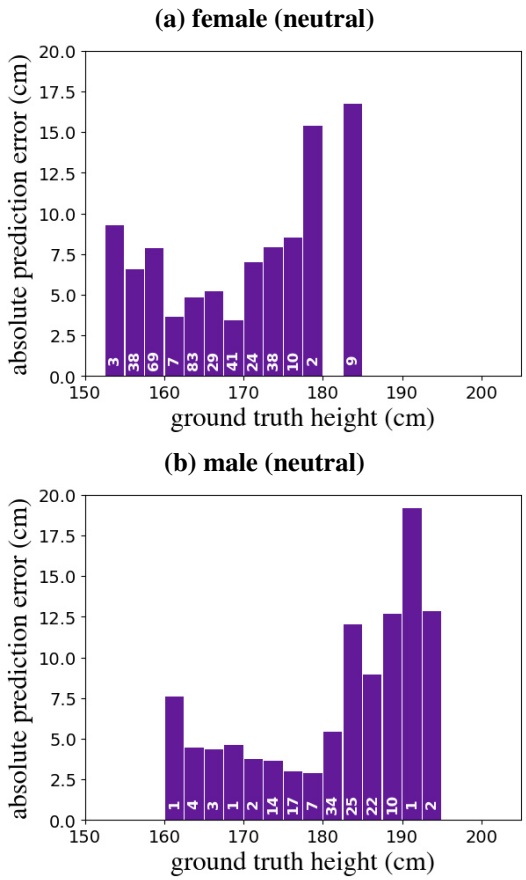

Figure 6. Shown is the absolute error between the estimated and ground-truth heights for the in-the-wild dataset, separated by gender. The value on each bar indicates the number of samples in each height range. Errors increase as the ground-truth height deviates from the mean adult height of 161 cm (female) and 175 cm (male) because the assumed IPD for scale disambiguation is height dependent. See also Figure 7.

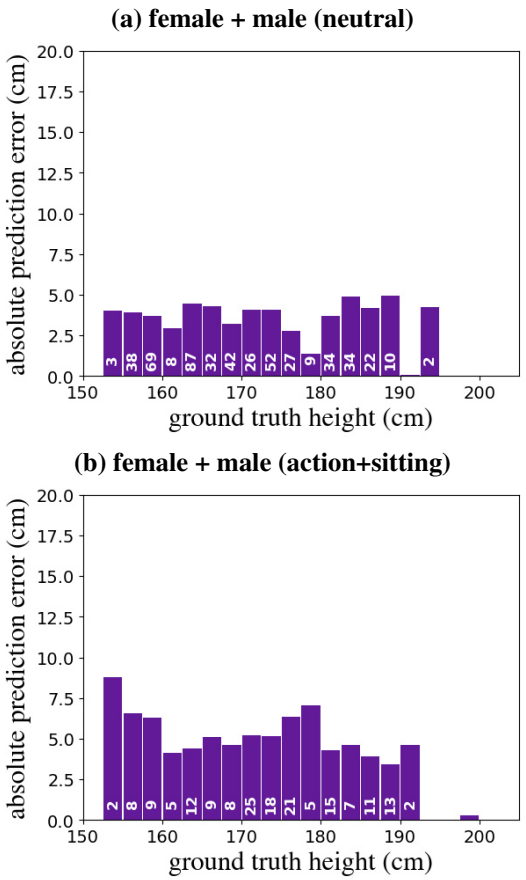

Figure 7. Shown is the absolute error between the estimated and ground-truth heights for the in-the-wild dataset, separated by pose. The value on each bar indicates the number of samples in each height range. Unlike Figure 6, the assumed IPD is scaled to be height dependent, significantly reducing errors.

3D pose estimation.

Shown in Figure 8 is the result of using the single-view metrology method of Criminisi et. al, with a known height of the beam supporting the roof, to estimate height [9]. This yields an estimated height of 178.6 cm relative to the ground-truth height of 180 cm. The authors in [9] hypothesize the slight under reporting of the height is due to the fact that the person is leaning on their right leg. Shown in the bottom panel of Figure 8 is a fitted 3D model, capturing the slight bend in the knees.

We start with the assumption that, as shown in Figure 8 (middle panel), the head to right foot distance in the fitted model is the previously estimated 178.6 cm. Then, by directly measuring the height of the model reposed into a neutral position, Section 3, height is estimated to be 179.9 cm, only 0.1 cm less than ground-truth.

This, obviously, is only one example and further study is required to determine by how much 3D pose estimation will improve standard single-view metrology across body poses and other factors. This example, however, nicely illustrates how even a slight deviation from a perfectly upright, neutral pose can impact height estimation, further emphasizing the importance of 3D pose estimation.

5. Discussion

We propose the use of 3D pose estimation to improve the accuracy with which height and weight can be forensically measured from a single, reference-free image.

When the estimated 3D model scale can be reliably determined, we find, in large-scale simulations, height can be accurately measured for a wide range of body poses and camera angles. The scale disambiguation used in these simulations (CPD), however, is not feasible in real-world scenarios. A feasible, but less accurate, scale disambiguation based on an average inter-pupillary distance (IPD) can lead

Figure 8. An example of augmenting single-view metrology with 3D pose estimation: (top) An original image analyzed in Criminisi et. al [9], where the person is 180 cm; (middle) the result of their analysis with a predicted height of 178.6 cm (middle) – the authors hypothesize that this under prediction is probably due to fact the person is leaning on their right leg; and (bottom) the result of our 3D model fitting superimposed atop the image – note how the 3D model captures the slight bend in the knees. Combining this 3D model with the known height of the beam yields an estimated height of 179.9 cm, only 0.1 cm less than ground-truth.

to accurate height estimations but for a more narrow range of body poses and camera angles. This general pattern of results generalizes to real-world scenarios, where we find, under limited conditions, height estimation can be reasonably accurate.

We hypothesize that a measured IPD, as opposed to an average IPD, from a known suspect will lead to more accurate height estimates. This approach, however, is only possible in certain real-world forensic scenarios and requires further validation.

We conclude that the 3D pose estimation is necessary, but not sufficient, to achieving accurate estimates of height and weight under a broad range of scene and imaging conditions. Resolving the scale ambiguity in a single, reference-free image remains challenging. When additional information is available in the scene (e.g., the known height of a doorway or floor tiles), the absolute scale can be determined more accurately. We hypothesize, therefore, that the addition of 3D pose estimation to existing techniques should lead to more accurate estimates of height. Beyond the one example provided above that combines 3D pose estimation with single-view metrology, further studies are required to better integrate these techniques and understand the reliability and accuracy of this approach.

In contrast to height estimation, weight (volume) estimation is no better than guessing a gender-specific average weight. We hypothesize this is because the 3D modeling focuses only on the 2D skeletal keypoints and does not consider the overall body shape. For non-average body shapes, therefore, the 3D body shape is not well modeled, even when the 3D pose is. As a result, weight cannot yet be accurately determined. We are investigating how appearance-based techniques can be incorporated into the 3D body pose estimation with the hope this will yield more accurate weight estimation. This will also require thought as to how to transform a 3D volume to metric weight.

Unlike many computer vision tasks where it is not possible or desirable to have a human in the loop, forensic analysts can manually intervene in an analysis by, for example, annotating known object sizes and manually refining the fitted 3D model. It is critical, however, to ensure that such manual interventions do not introduce bias.

Even the seemingly simple and most basic forensic analysis of measuring height and weight is riddled with complexities: apparent height is impacted by body pose and camera angle, footwear and headwear, and physiological changes throughout the day; weight is impacted by clothing and physiological fluctuations; and estimating metric scale from a single reference-free image is challenging. While 3D pose estimation helps to contend with some of these complexities, many remain. Caution, therefore, should be taken when making these forensic measurements. It is our hope that further advances in body pose and shape estimation and scale disambiguation, along with large-scale studies of accuracy and reliability, will continue to improve the state of digital forensic identification.

Acknowledgements

We thank Angjoo Kanazawa for many helpful discussions and advice.

References

[1] NHANES questionnaires, datasets, and related documentation. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015. Accessed: 2020-11-20. 2

[2] Ivo Alberink and Annabel Bolck. Obtaining confidence intervals and likelihood ratios for body height estimations in images. *Forensic Science International*, 177(2-3):228–237, 2008.

[3] Olasimbo Ayodeji Arigbabu, Sharifah Mumtazah Syed Ahmad, Wan Azizun Wan Adnan, Salman Yussof, Vahab Iranmanesh, and Fahad Layth Malallah. Estimating body related soft biometric traits in video frames. *The Scientific World Journal*, 2014, 2014. 2

[4] C. BenAbdelkader and Y. Yacoob. Statistical body height estimation from a single image. In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–7, 2008. 2

[5] Didier Bieler, Semih Günel, Pascal Fua, and Helge Rhodin. Gravity as a reference for estimating a person's height from video. *International Conference on Computer Vision*, pages 8568–8576, 2019. 2

[6] Zhe Cao, T. Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1302–1310, 2017. 3

[7] Rama Chellappa and Pavan Turaga. Recent advances in age and height estimation from still images and video. In *Face and Gesture*, pages 91–96, 2011.

[8] A. Criminisi, I. Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision*, 40:123–148, 2004. 2

[9] Antonio Criminisi, Andrew Zisserman, Luc J. Van Gool, Simon K. Bramble, and David Compton. New approach to obtain height measurements from video. In Kathleen Higgins, editor, *Investigation and Forensic Science Technologies*, volume 3576, pages 227 – 238. International Society for Optics and Photonics, SPIE, 1999. 2, 7, 8

[10] Antitza Dantcheva, Francois Bremond, and Piotr Bilinski. Show me your face and I will tell you your height, weight and body mass index. In *International Conference on Pattern Recognition*, pages 3555–3560. IEEE, 2018.

[11] Neil A Dodgson. Variation and extrema of human interpupillary distance. In *Stereoscopic Displays and Virtual Reality Systems XI*, volume 5291, pages 36–46. International Society for Optics and Photonics, 2004. 3

[12] Harry T Edwards. Ten years after the National Academy of Sciences' landmark report on strengthening forensic science in the United States: A path forward–where are we? *Available at SSRN 3379373*, 2019. 1

[13] Dale A Gerke, Jean-Michel Brismée, Phillip S Sizer, Gregory S Dedrick, and C Roger James. Change in spine height measurements following sustained mid-range and end-range flexion of the lumbar spine. *Applied ergonomics*, 42(2):331–336, 2011. 2

[14] Semih Günel, H. Rhodin, and P. Fua. What face and body shapes can tell us about height. *IEEE/CVF International Conference on Computer Vision Workshop*, pages 1819–1827, 2019. 4

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, Oct. 2019. 2

[16] Dong-Seok Lee, Jong-Soo Kim, Seok Chan Jeong, and Soon-Kak Kwon. Human height estimation by color deep learning and depth 3D conversion. *Applied Sciences*, 10(16):5531, 2020.

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 3

[18] Albrecht Meydenbauer. Die photometrographie. *Wochenblatt des Architektenvereins zu Berlin*, (14):125–126, 1867. 2

[19] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:2262–2275, 2010. 3

[20] National Research Council Committee on Identifying the Needs of the Forensic Sciences Community. *Strengthening forensic science in the United States: A path forward*. National Academies Press, 2009. 1

[21] Tam V Nguyen, Jiashi Feng, and Shuicheng Yan. Seeing human weight from a single RBG-D image. *Journal of Computer Science and Technology*, 29(5):777–784, 2014. 2

[22] Sophie Nightingale and Hany Farid. Assessing the reliability of clothing-based forensic identification. *Proceedings of the National Academy of Science*, 117(20):5176–5183, 2020. 1

[23] Angela Olver, Helen Guryn, and Eugene Liscio. The effects of camera resolution and distance on suspect height analysis using PhotoModeler. *Forensic Science International*, page 110601, 2020.

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 3, 4

[25] Carmelo Velardo and Jean-Luc Dugelay. Weight estimation from visual body appearance. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2010. 2

[26] Paul R Wolf and Bon A Dewitt. *Elements of photogrammetry: with applications in GIS*, volume 3. McGraw-Hill New York, 2000. 2

[27] Rui Zhu, X. Yang, Yannick Hold-Geoffroy, Federico Perazzi, J. Eisenmann, Kalyan Sunkavalli, and M. Chandraker. Single view metrology in the wild. *arXiv*, abs/2007.09529, 2020. 2