Estimation of Body Measurements from Personal Pictures

John AOUSSOU
Lead ML Researcher at Sapeet Inc.
2018-2019

1 Introduction and Problem Statement

This case study outlines research conducted around 2018 into the feasibility of obtaining human body measurements using Computer Vision techniques applied to full-sized personal photographs (e.g., full-body shots) taken by consumers, using smartphones. The primary objectives were twofold:

1. To enable individuals to determine their body measurements from multiple full-sized pictures (requiring different angles), facilitating accurate size selection for online clothing purchases.
2. To generate a personalized digital avatar resembling the individual based on these estimated measurements.

The project operated under specific initial constraints:

- **Input Data Limitation:** The only available inputs were the multiple photographs of the subject, the subject's self-reported height, and the approximate distance of the camera from the subject for each photo (assuming no camera tilt). No other calibration data or specialized equipment could be assumed.
- **Methodology Requirement:** Utilization of at least one deep learning model was mandatory.
- **Integration Constraint:** Integration with the company's existing avatar model (based on Blender) was required.
- **Training Data Constraint:** Only publicly available datasets could be used for model training.

2 Methodology and Initial Results

The initial strategy, adhering to the constraint of using only publicly available data sources, involved evaluating a model trained as follows:

1. **Dataset Creation:** A large dataset of human silhouettes was created synthetically using the company's Blender-based avatar model. To enhance robustness, reasonable variations in human body shapes and dimensions were introduced, along with variations in the virtual camera angles used to capture the avatar images (simulating multiple viewpoints).
2. **Ground Truth Generation:** A custom measurement module was developed to precisely measure the body dimensions of the synthetic avatars. Measurements were based on predetermined key points on the avatar mesh, providing accurate ground truth labels for training.
3. **Model Training:** The model architecture described in the paper *HS-Nets: Estimating Human Body Shape from Silhouettes with Convolutional Neural Networks* was reproduced from scratch

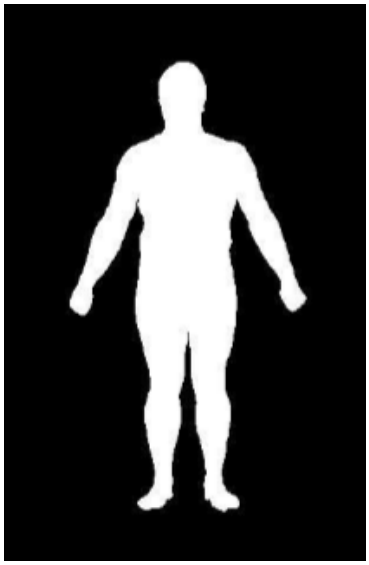

Figure 1: Example of a silhouette input, illustrating the type extracted from photographs and used for body shape estimation (Image source: Dibra et al., 2016).

and adapted to suit the synthetic dataset format. This adapted model was then trained on our dataset, aiming to predict body shape parameters directly from the input silhouettes.

4. **Real-world Application Pipeline:** For application on real images, the intended pipeline involved capturing multiple full-sized photographs of the user from different angles (e.g., front, 30 degrees, 60 degrees). Silhouettes would be extracted from these photographs using a suitable image segmentation model. These multiple silhouettes would then be fed into the trained HS-Nets model to estimate body measurements, using the provided height and camera distance to attempt scaling.

Initial Evaluation: The model achieved very good performance on the test set drawn from the synthetic dataset itself, closely matching the ground truth measurements. For this synthetic test set, the error would typically be less than $1 \sim 2$ cm. When applied to a small in-house test set of real images (approx. 10 subjects), the generated digital avatars often captured reasonable overall body shape and proportions. This visual plausibility occurred because the model learned relative shape from silhouettes effectively, and the avatar generation primarily relied on these proportions, making it less sensitive to precise scaling inaccuracies inherent in converting shape to absolute metric dimensions using only approximate height and distance.

However, the *absolute measurements* derived for these real subjects were significantly poorer and unreliable compared to the synthetic results. For the real case test, errors could sometimes be as large as 5 cm or more off. This quantitative failure was attributed to the domain gap between synthetic training data and uncontrolled real-world images. It was suspected that the model had effectively memorized the limited variations in the synthetic data rather than learning generalizable features needed for accurate measurement on real photos. This lack of reliable measurement accuracy prompted a revision of the approach.

3 Further Results and Discussion

Given the poor generalization observed with models trained purely on synthetic data and the limitations of publicly available real-world datasets suitable for this task at the time, the methodology was revised:

1. **Dataset Acquisition (Relaxing Initial Constraint):** To incorporate more realistic training data, the initial constraint limiting data sources to publicly available datasets was relaxed. The decision

was made to purchase the licensed CAESAR (Civilian American and European Surface Anthropometry Resource) dataset, which contains 3D scans of real human subjects.

2. **New Challenges:** The use of real scan data introduced new complexities:
    - **Key Point Ambiguity:** Unlike the synthetic avatars generated from a consistent model, the real human scans lacked strictly defined, consistently located key points. Establishing reliable landmarks for measurements became challenging. For example, attempting to measure the inseam required identifying the approximate vertex where the two legs connect, a location that varied between scans and could be ambiguous.
    - **Scan Artefacts:** Some scans contained spurious vertices or mesh irregularities, hindering the reliable application of the custom measurement module developed for the clean synthetic avatars.

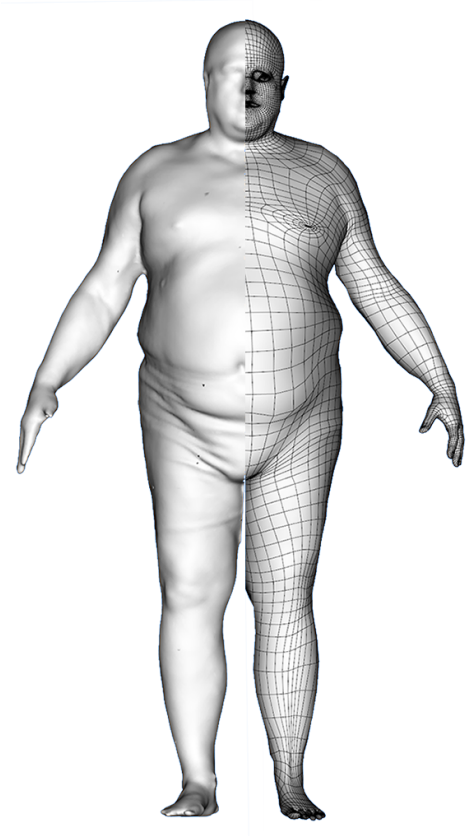

Figure 2: Example of a 3D body scan from the CAESAR dataset, representing the real-world anthropometric data used in the revised approach. (Image source: humanshape.org/CAESAR)

Transitioning to the CAESAR dataset for training did not yield the expected improvement in performance on the in-house real test set. The results remained inconsistent and often inaccurate, mirroring the difficulties seen with the initial model when applied to real images. The issues highlighted during the initial evaluationspecifically the difficulty in obtaining accurate absolute measurements despite plausible shape representationpersisted. This suggested that the core challenges might not solely lie in the choice or realism of the training data (synthetic vs. real scans), but also in fundamental physical and definition-related issues inherent to the task of measurement from multiple, uncalibrated 2D images taken in uncontrolled settings, even with height and approximate distance provided.

4 Analysis of Persistent Challenges

Further investigation revealed several underlying reasons for the consistently low performance on real-world data, largely independent of the specific training dataset:

1. **Dependence on Scale Information:** Even with multiple views, the subject's height, and approximate camera distance, accurately determining absolute dimensions remained challenging. Deep learning models trained on silhouettes primarily learn body *ratios* or relative shape. Converting these into actual measurements (e.g., centimeters or inches) requires precise knowledge of the subject's height *and* the exact camera positions and calibration parameters relative to the subject.
    - While multiple views *can* theoretically help resolve 3D geometry (e.g., via Structure from Motion), doing so accurately from uncalibrated smartphone photos taken by untrained users is notoriously difficult. The provided height and approximate distance were insufficient to overcome ambiguities and errors in pose, camera intrinsics, and segmentation needed for metric scale reconstruction.
    - Small errors in the reported height, the estimated camera distance, the assumed lack of camera tilt, or slight deviations in user positioning could introduce significant inaccuracies in the final calculated dimensions.
2. **Ambiguity in Definitions:** Defining "height" or the precise location of anatomical key points on a real, non-rigid human body remains inherently ambiguous and subject to variations in posture and measurement protocol. This affects ground truth definition and evaluation.
3. **Image Quality and Segmentation (Per Image):** Issues apply to each image captured:
    - **Segmentation Accuracy:** The initial silhouette extraction step for *each view* is critical. Inaccuracies in segmentation (especially with less robust models available at the time, or in challenging conditions) directly propagate errors.
    - **Image Resolution:** Resolution limits apply to each photo. At typical distances, a difference of even a single pixel in a silhouette boundary could translate to significant measurement error.
    - **Lens Distortion:** Each smartphone camera lens introduces geometric distortions that affect the perceived shape if not accurately corrected (requiring calibration data often unavailable).
    - **Lighting and Pose Consistency:** Ensuring consistent lighting and precise, repeatable poses across multiple user-taken photos is very difficult.
4. **Human Body Diversity and Definition:** The sheer diversity of human bodies makes defining and consistently measuring specific dimensions challenging, regardless of the number of views.

5 Conclusion

This 2018 study explored estimating body measurements from user photos using only height and approximate distance. Initial deep learning models trained on synthetic data yielded plausible avatar shapes but failed on real images for accurate measurements. Using the licensed CAESAR dataset similarly proved insufficient. Key challenges included the domain gap, obtaining reliable scale from uncalibrated, user-captured photos (highly sensitive to errors in inputs, pose, and segmentation), and anatomical ambiguities. The conclusion circa 2018 was that purely vision-based methods with such limited data were impractical for high-accuracy sizing, suggesting hybrid or statistical models were likely more viable at the time.