Body Talk: Crowdshaping Realistic 3D Avatars with Words

Stephan Streuber[1] M. Alejandra Quiros-Ramirez[1] Matthew Q. Hill[2]
Carina A. Hahn[2] Silvia Zuffi[3,†] Alice O'Toole[2] Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of Texas at Dallas, [3]ITC-CNR, Milan, Italy

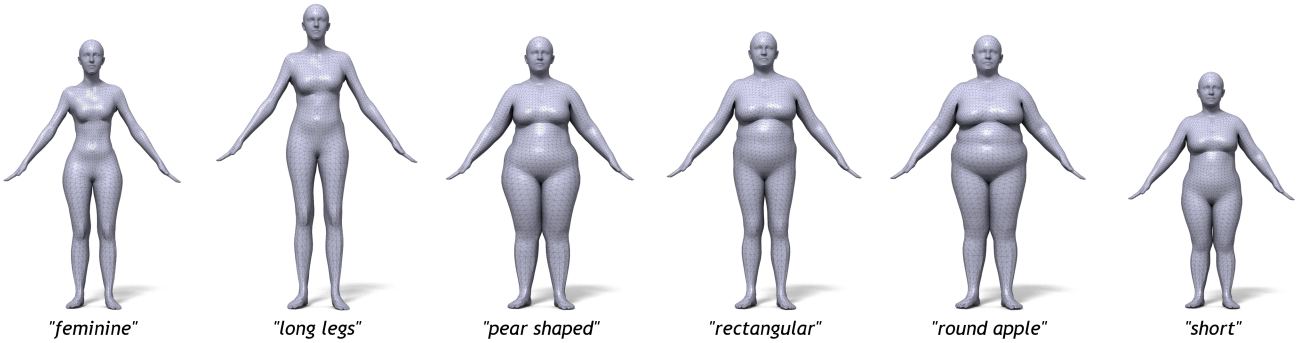

Figure 1: Prototypical body shapes. *We generate random 3D body shapes, render them as images, and then crowdsource ratings of the images using words that describe shape. We learn a model of how 3D shape and linguistic descriptions of shape are related. Shown are the most likely body shapes, conditioned on the words below them. The ratings of "the crowd" suggest that we share an understanding of the 3D meaning of these shape attributes.*

Abstract

Realistic, metrically accurate, 3D human avatars are useful for games, shopping, virtual reality, and health applications. Such avatars are not in wide use because solutions for creating them from high-end scanners, low-cost range cameras, and tailoring measurements all have limitations. Here we propose a simple solution and show that it is surprisingly accurate. We use crowdsourcing to generate attribute ratings of 3D body shapes corresponding to standard linguistic descriptions of 3D shape. We then learn a linear function relating these ratings to 3D human shape parameters. Given an image of a new body, we again turn to the crowd for ratings of the body shape. The collection of linguistic ratings of a photograph provides remarkably strong constraints on the metric 3D shape. We call the process *crowdshaping* and show that our *Body Talk* system produces shapes that are perceptually indistinguishable from bodies created from high-resolution scans and that the metric accuracy is sufficient for many tasks. This makes body "scanning" practical without a scanner, opening up new applications including database search, visualization, and extracting avatars from books.

Keywords: Human body modeling, body shape, 3D shape, avatars, crowdsourcing, perception, anthropometry

Concepts: •**Computing methodologies** → **Shape modeling;** *Perception;*

SIGGRAPH '16 Technical Paper, July 24-28, 2016, Anaheim, CA,
ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925981

1 Introduction

> *He was of medium height, solidly built, wide in the shoulders, thick in the neck, with a jovial heavy-jawed red face . . .*
>
> Dashiell Hammett, The Maltese Falcon [1929]

Language can create a visual representation of a character in the mind of the reader. That is, words conjure bodies. But do the same words create the same representation in all of us? To what degree do we have a shared understanding of language as it relates to 3D shape? Can we communicate 3D shape to each other (and to a computer) precisely with words alone? Does one need an expensive or complex 3D scanner to obtain an accurate model of body shape?

We attempt to answer these questions in the context of human shape. We hypothesize that 1) people have a shared understanding of shape that is reflected in our use of language; 2) the collective judgement of shape attributes by "the crowd" contains a robust signal about body shape; and 3) correlations in the ratings of shape attributes, and their relationship to shape statistics, provide sufficient constraints from which to estimate metrically accurate 3D shape.

Bodies and their shape are important for communication, recognition of identity, and conveying emotion. Shape further is an indicator of gender, age, health, and fitness. Arguably, the human body is the object with which we are most familiar and, not surprisingly, our language for communicating body shape is rich. These properties make human bodies a good test case for modeling the relationship between language and shape.

We also focus on the body because the demand for realistic 3D digital avatars is expanding with applications in games, virtual reality, on-line shopping, and visual effects. Realistic 3D bodies can be created from high-end scanners, low-cost range cameras, and

[†]This work was performed while SZ was at the MPI for Intelligent Systems and the Bernstein Center for Computational Neuroscience.

tailoring measurements. High-end scanners (laser, structured light, stereo) produce realistic avatars [Allen et al. 2003; Allen et al. 2006; Anguelov et al. 2005; Hasler et al. 2009; Hirshberg et al. 2012; Loper et al. 2015] but are costly and not widely available. There are many methods that extract avatars from range cameras [Bogo et al. 2015; Li et al. 2013; Shapiro et al. 2014; Weiss et al. 2011], typically with lower quality, but even these sensors are not yet widely available. Traditional tailoring measurements can be taken by anyone with a tape measure, and have been used to create avatars [Allen et al. 2003; Hasler et al. 2009; Seo and Magnenat-Thalmann 2003; Seo et al. 2003], but the approach is error prone. Naive subjects exhibit significant variance in measurements [Gorber et al. 2007; Spencer et al. 2002] as do even experienced anthropometrists, using a well defined protocol [Gordon et al. 1989]. For the wide use of realistic 3D avatars in shopping, games, fitness, etc., a simple, easy to use, low-technology, and fun body creation solution is required. In this paper we propose a novel method for estimating perceptually and metrically accurate 3D geometry of human bodies in an intuitive and technologically inexpensive way.

Our *Body Talk* system requires only one photograph of a person and 15 people who rate the body shape in the photo using 30 words (or fewer). In a training phase, we first crowdsource linguistic descriptions of synthetic bodies and perform linear regression to model the mapping from linguistic body descriptions to the geometric description of body shape. This model allows us to generate 3D digital bodies from verbal descriptions and vice versa. We call this process *crowdshaping*. While we can create a body with ratings from a single person, we show that accuracy increases as more people rate the body shape. A collective view of body shape emerges and the varied linguistic descriptions constrain body shape with surprising metric accuracy.

By relating the crowd-sourced attributes to body shapes we can explore this collective linguistic understanding of shape by creating and visualizing bodies that correspond to attribute dimensions (cf. [Blanz and Vetter 1999]). We do this by conditioning our model on one or more shape attributes and generating the body shape most consistent with these. Figure 1 shows several examples of body shapes generated from our model in this way.

The geometric body space underlying Body Talk is provided by the SMPL model [Loper et al. 2015]. SMPL is a vertex-based 3D model that accurately represents a wide variety of body shapes and poses in a low dimensional parameter space. SMPL uses principal component analysis (PCA) to learn a pose-independent Euclidean representation of body shape. We use the first eight principal components (PCs) of the publicly available shape model.

In contrast to recent work that uses comparative or relative ratings [Chaudhuri et al. 2013; Liu et al. 2015; Lun et al. 2015] we directly crowdsource scalar ratings of body shapes on a 5-point scale. Using only these attribute ratings, Body Talk produces 3D crowdshaped models with an average Euclidean vertex-to-vertex error of about 9 mm, compared with ground truth meshes. We can include self-reported (noisy) height and weight in the model, reducing the error to about 8 mm. We also quantitatively evaluate the prediction of anthropometric measurements and find that average absolute error in linear measurements is 6.68 mm and 8.90 mm for circumferential measurements. These errors are low enough to be useful for clothing sizing applications. We further show that the crowdshaped bodies are perceptually accurate. In fact, we found that they are perceptually indistinguishable from bodies constructed from high-resolution scans. Thus, our model provides a good tradeoff between simplicity and geometric accuracy without compromising the perceptual fidelity of the bodies.

We show several novel applications of Body Talk including crowdsourcing body shapes of celebrities from photographs, creating characters from books, and creating 3D bodies corresponding to the classical somatotypes (Endomorph, Ectomorph, Mesomorph). We also invert our model to compute semantic attributes for all the bodies in the CAESAR dataset [Robinette et al. 2002], enabling semantic queries for bodies with particular shapes.

In summary, we make several unique contributions relative to the prior art: 1) We create bodies from generic adjectives and adjectival phrases describing shape attributes, rather than words related to metric properties. 2) We do this using scalar ratings rather than relative judgements. 3) We show that crowd-sourced linguistic ratings provide a robust signal about body shape. 4) We show that, from these ratings, we can recover *perceptually and metrically accurate* body shapes (crowdshaping). 5) We demonstrate several novel applications that exploit crowd-sourced body shape. 6) We provide a website (http://bodytalk.is.tue.mpg.de/) that lets users create avatars with linguistic sliders, understand relationships between words and body shape, and download meshes for research purposes.

2 Related work

Body models. Since the introduction of the CAESAR dataset [Robinette et al. 2002], there has been significant work on learning statistical models of 3D human body shape. Allen et al. [2003; 2004] were the first to learn a statistical body model and relate it to measurements. They align a template to 250 people in the CAESAR dataset and, without factoring out pose variation, perform PCA. They then learn a linear function relating anthropometric measurements to the PC coefficients. They show how to create and edit bodies using "sliders" that vary the coefficients along the principal component directions.

Seo and Magnet-Thalmann [2003] take a similar approach to relating shape to measurements using a non-linear mapping (radial basis functions). More relevant to our problem, Seo et al. [2003] suggest trying to relate attributes such as "hourglass" or "pear/apple" to body shape but conclude that these are too abstract to measure. Consequently, they focus instead on relating shape to numerical quantities that they can estimate: hip-to-waist ratio (HWR), fat percentage (estimated from other measurements), and height.

Hasler et al. [2009] describe a method to rotate a body shape PCA space to create a "semantic model basis." They represent body shape using triangle deformations [Sumner 2005]. In addition to defining morphing vectors based on measurements like height and weight they model more semantic directions like "muscularity." They argue that it is hard for humans to rate muscularity so they present subjects with pairs of bodies and ask them to determine which is more muscular. They convert these pairwise judgements into a semantic direction that can be used to change body shape [Jain et al. 2010]. Also Zhou et al. [2010] use "semantic" attributes to edit body shape in images but their "semantic" attributes are metric quantities like height, weight, and girths. We refer to these as "metric" rather than semantic because they are directly measurable on the mesh or in the world. We show how to recover body shape without metric attributes using semantic attributes like "hourglass" and "feminine".

The above do not show how to create accurate bodies directly from linguistic descriptions and do not use crowdsourcing. In contrast to previous methods we show that it is not necessary to resort to pairwise judgements and that the collective ratings of the crowd carry significant metric information.

Face models. The ability for police sketch artists to create images from verbal descriptions is a proof of concept that humans

have a shared visual vocabulary that can be used to transmit an image to another person. Previous work in computer vision attempts to relate verbal descriptions to images of faces (see [Klare et al. 2014] for a recent example) without modeling 3D shape.

Blanz and Vetter [1999] pioneered the relationship between attributes and facial shape. They take a similar approach to ours but work only with face shape and appearance. For each exemplar face, they define a vector representing how it differs from the mean face shape and appearance. Then for different attributes like "hooked nose," "distinctiveness," and gender, the experimenters assign weights to the exemplars. From this they construct 1D vectors representing variation along that attribute direction in shape and appearance space. They do not use crowd-sourced data and do not model correlations in the attributes ratings. While they demonstrate face shape editing using attributes, they do not show that metrically accurate faces can be constructed from words alone.

Using the 3D model of Blanz and Vetter [1999], O'Toole et al. [1999] ask subjects to rate faces for age and attractiveness on a 5-point scale. They relate the ratings to 3D shape to create new faces with varying age and attractiveness. Face images have also been rated for semantic attributes such as trustworthiness and the ratings have been used to morph 2D face images to make them more or less trustworthy [Little et al. 2012].

Reid et al. [2013; 2014] describe "soft biometrics", which relate human verbal descriptions to physical or behavioral traits. Their focus is on creating soft biometric signatures for recognition of individuals rather than metric reconstruction of shape. Like others, they focus on comparative ratings.

Computer graphics and vision. There is recent work in computer graphics on relating physical properties to perceptually salient semantic attributes. The motivation is typically to provide more intuitive controls for animating complex physical processes. For example, Sigal et al. [2015] learn a mapping between semantic ratings of simulations and cloth parameters, enabling animators to control cloth simulations in a more natural way. Similar approaches have been taken in modeling fonts [O'Donovan et al. 2014], BRDFs [Matusik et al. 2003], and crowd simulations [Guy et al. 2011].

Similar in spirit to our work, Yumer et al. [2015] relate semantic attributes to 3D object shape. Like Hasler et al. [2009], they use pairwise judgements; two objects are shown and the subjects rate their similarity along different semantic axes. In contrast to our body shapes, which are in correspondence, they deal with the challenging case of objects where correspondence cannot be easily established. They do not, however, create shapes from crowd-sourced data or evaluate metric accuracy.

Lun et al. [2015] use crowd-sourced ratings to learn a measure of style similarity that attempts to match human judgements of style. They deal with man-made objects and focus on the similarity of object parts. They use pairwise comparisons of style and also convert this into a distance function. Liu et al. [2015] address style compatibility of 3D furniture using triplets of relative style compatibility ratings. Using style-aware shape retrieval, the approach generates discrete suggestions of scenes based on the learned compatibility of object shapes. Xia et al. [2015], describe a system for style transfer in the context of motion capture animation. They mocap people doing specific actions in various styles, giving them pre-labeled semantic attributes; hence they do not use crowd-sourced ratings. Chaudhuri et al. [2013] describe a system for creating models from parts with different semantic attributes. They use relative attribute ratings to create a ranking of *discrete* sets of parts along semantic axes. They do not address metric accuracy nor do they create shapes from language alone or from verbal description of photos.

Talton et al. [2009] describe a method for "exploratory modeling" and show examples with human bodies. They use the PCA shape space of Allen et al. [2003] and define an interface that shows 124 exemplar bodies. Users create new bodies by interacting with this space and the space itself is updated with the models people create. Each body is created by one person; the shapes are not crowd-sourced. This work does not address metric accuracy of the created shapes and does not relate these shapes to words.

Extracting semantic attributes from images is widely studied in computer vision but a thorough review is outside the scope of this paper. For example, Parikh and Grauman [2011] learn how to recognize attributes from image data. Such work does not address our synthesis problem and does not focus on 3D shape. Pons-Moll et al. [2014] relate 3D articulated human pose to semantic attributes but do not consider body shape.

Psychology. Sheldon [1940] proposes "a three-dimensional system for the description of the human physique." This system for classifying body shapes, or *somatotyping*, is based on photographs and measurements, resulting in three categories of bodies: Endomorph, Mesomorph, and Ectomorph [Carter and Heath 1990]. Note that Sheldon does not use standard adjectival words for shape. Instead he creates specific, novel, shape categories and then classifies bodies into these categories. The theory has been generally discredited, not because of the shape categorization but because it tries to relate these body shapes to "temperament." We do something quite different in that we explore how our shared vocabulary for shape description is related to 3D shape.

The most related modern work in psychology is [Hill et al. 2015]. In that work, humans rated photographs of people using shape adjectives. For each word, subjects rated photos with one of three values: *does not apply*, *applies somewhat*, or *applies perfectly to the body.* The authors then computed a space of word ratings using correspondence analysis (CA) [Greenacre 2007]. CA was applied to ratings of photos from 164 females, using only the ratings that "applied perfectly." This created a space of verbal descriptions of bodies.

Separately, they took a PCA space of body shapes represented using deformation gradients [Anguelov et al. 2005; Hirshberg et al. 2012; Sumner 2005]. They noted a qualitative similarity between the first 5 axes of the shape and linguistic spaces (though the ordering of the components was different) and they manually reordered the axes so that they appeared to capture similar properties. They showed that the reordered axes of the two spaces were highly correlated. They synthesized 3D bodies by taking the coefficients from the language space and using these exact values in the reordered shape space. These recovered 3D body models were then described by human subjects, following the same procedure applied to the photographs. The descriptions of the photographs and the synthesized bodies were strongly correlated. The findings demonstrate that human language descriptions capture perceptually salient variations in human body shape and that these descriptions can be transferred to a body shape space to create 3D bodies.

We go beyond that work in several important ways. First, we use crowdsourcing and show that it is reliable. Second, we have people rate 3D bodies and directly learn a linear mapping from ratings to shape parameters. We use the same shape adjectives here but have people rate 3D shapes using a 5-point Likert scale. Third, we use a different body model with shape represented as principal components in Euclidean vertex space [Loper et al. 2015; SMP 2015]. This captures more body shape variation than the model used in [Hill et al. 2015]. Fourth, we perform our analysis with both male and female bodies. Fifth, we learn the correlations in the ratings, allowing us to condition on particular ratings to infer the rest. This al-

Average	Big	Broad Shoulders	Built
Curvy	Feminine	Fit	Heavyset
Lean	Long Legs	Long Torso	Long
Masculine	Muscular	Pear Shaped	Petite
Proportioned	Rectangular	Round Apple	Short Legs
Short Torso	Short	Skinny	Small
Stocky	Sturdy	Tall	*Attractive*
Sexy	*Hourglass*		

Table 1: Shape attributes. *Linguistic descriptors used in rating body shape. See text.*

lows us to construct bodies with particular shape properties. Sixth, we demonstrate the metric accuracy of bodies created from words in multiple ways. Finally, we show how this method can be used for a variety of novel applications.

3 Methods

Our goal is to learn a mapping between a linguistic body space and a geometric body space.

The geometric body space is provided by the identity shape component of SMPL [Loper et al. 2015]. The body is represented by a 3D template mesh with 6890 vertices. The template mesh is registered to high-resolution body scans in the CAESAR dataset [Robinette et al. 2002], resulting in 1700 registred meshes for males and 2100 for females. Variations in pose are removed to create a shape training dataset for PCA. A body shape is represented by a mean mesh and a linear combination of deviations from the mean along the principal component directions. In this study we use the first 8 principal shape components downloaded from [SMP 2015]. These account for 96.56% of the identity-related body shape deformations in the shape training dataset.

The linguistic space is represented in terms of 30 body descriptor words (e.g. curvy, fit, heavyset, round-apple) as suggested in [Hill et al. 2015]. They collected these words by asking human participants to tag photographs of female bodies using adjectives. Their linguistic space represents a vector of the 27 most frequently used adjectives. Table 1 shows the complete list, including three additional words in italics that were not used in [Hill et al. 2015]. Using Amazon Mechanical Turk (MTurk) we collected ratings of bodies with respect to these words. Each body was rated by at least 15 raters and there were a total of 256 raters in the study.

3.1 Data Collection

We used the identity component of SMPL to generate 128 synthetic female and 128 synthetic male bodies in a neutral pose by randomly sampling the first 8 principal shape directions. We tried several different ways of generating training bodies including, sampling uniformly along each PC direction, taking bodies at fixed distances from the mean, and sampling bodies at random from the CAESAR dataset. We found that sampling bodies from a Gaussian distribution, using the variances given by PCA, gave the best results.

The resulting 128 female and 128 male meshes represent synthetic bodies that express the global features of expected body shape variation in the female and male population as captured by CAESAR. We set the pose of the bodies to the mean pose of women and men in the CAESAR dataset respectively. Figure 2 shows an example stimulus image of a random female body. Each synthetic body was rendered in Maya using the same camera model for all subjects. The viewing direction and lighting were the same for all stimuli; bodies were only shown with a frontal orientation. The

Figure 2: HIT. *Example of the task shown to MTurk raters.*

feet were always in the same vertical location, meaning that the height of the person in the image conveyed relative information about the person's 3D height. One can see examples and obtain the meshes for research purposes from https://ps.is.tuebingen.mpg.de/research_projects/bodies-from-words.

In order to establish a relationship between the geometric shape space and the linguistic shape space, 256 MTurk users rated images of the synthetic bodies using 30 descriptive words; 128 rated females and 128 rated males. While viewing a synthetic body on the screen, participants rated the body shape according to each word on a 5-point Likert scale: [(1) does not apply at all, (2) does not apply, (3) average, (4) does apply, (5) completely applies]. While such scalar ratings are common in psychology, recent work in graphics has focused on relative ratings that compare attributes between pairs or triples of images. The argument is that relative ratings are easier for users to make. Such ratings are typically used in discrete classification tasks, where they work well. Our goal is to recover metric, scalar, properties of body shape and converting relative ratings into metric distances remains an open research problem [Kleindessner and von Luxburg 2015]. While the scalar ratings of individual raters are not reliable, we find that the combined scalar ratings of "the crowd" are sufficient for metric reconstruction.

Each human intelligent task (HIT) consisted of a qualification test, the rating of 15 synthetic bodies, and the rating of 2 extra bodies, which were used as catch trials. An example HIT is illustrated in Fig. 2. Each participant performed only one HIT. The catch trials consisted of the presentation of an extremely "skinny" or "big" body. Participants who did not rate the catch trials correctly were excluded from the experiment (approx. 10% of the participants). In order to ensure that participants understood the shape attribute words, we performed a language qualification test, which required participants to find the right synonyms for different adjectives. Only those participants who passed the qualification test were allowed to participate in the HIT. During the HIT, each rating task was displayed for at least 30 seconds, to make sure that the participants were not assigning ratings randomly without carefully considering the word descriptors. Since we aimed to test fluent English speakers we restricted the participants to those located in the US.

Participants accepted voluntarily to join the study and they were rewarded with 2.5 USD, corresponding roughly to a wage of 6.00

USD per hour. After each session, we collected demographic data (e.g. gender, age, nationality). The final collected dataset consists of at least 15 ratings for every word descriptor for each of the 256 synthetic bodies. The dataset is split into the ratings for the 128 female and 128 male synthetic bodies.

3.2 Training

Consider a single gender. Let the shape of body $i \in 1, \ldots, 128$ be a vector $\mathbf{y}_i = [\beta_1, \ldots, \beta_8]^T$ where the β's are the linear coefficients that represent body shape in the PCA space. Let the vector of ratings for each rater k and body i be a vector $[r_{1,i,k}, \ldots, r_{W,i,k}]^T$, where $W = 30$ words. The individual ratings are noisy and we found it useful to average the ratings for a body over the raters, giving 128 rating vectors that we denote $\mathbf{x}_i = [\bar{r}_{1,i}, \ldots, \bar{r}_{W,i}]^T$. We also tested median rating vectors with similar results.

Our observation matrix is then

$$X = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_{128}^T \end{bmatrix} \quad (1)$$

and the bodies are represented in $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_{128}]^T$ with one body per row. Assuming a linear relationship between ratings and shape coefficients, we solve for the regression coefficients B in

$$Y = XB + \epsilon \quad (2)$$

using least squares.

This defines our words-to-shape model (**w2s**). Given a new rating vector $\mathbf{x}$, we multiply by B to obtain the body shape coefficients $\mathbf{y}$, which define the shape in the SMPL PCA space.

Conditioning on ratings. Different shape descriptors like "skinny" and "petite" are correlated and it is important to model this.

Let $X' = X - \mu$ be the zero mean rating matrix, where we have subtracted the mean rating vector, μ. The covariance matrix $X'^T X'$ represents correlations in the ratings of different words and defines a multi-variate Gaussian distribution over the word ratings. This is useful because we can condition on one or more shape attributes (setting them to a constant) and generate the most likely ratings of the other words.

Consider the images in Fig. 1. To generate the body exemplified by a particular word, we compute the standard deviation of the ratings of that word. We then set the rating value for the word to its mean rating plus 4 times the standard deviation. We condition the Gaussian on this value and estimate the expected value of the remaining ratings, obtaining a synthetic rating vector. We then use the w2s model to generate the body shape for the predicted ratings.

Additional cues. As we show below, the w2s model is able to recover surprisingly accurate 3D body shapes. For some applications we might have more input data and may want higher accuracy. Specifically, most people roughly know their height and weight. We can include height, weight, or both in the rating vector as $\mathbf{x}_i = [h_i, w_i, \bar{r}_{1,i}, \ldots, \bar{r}_{W,i}]^T$, where h_i and w_i represent the (possibly wrong) height and weight of subject i. We augment the rating vectors in this way and train two additional models: "words and height" (**wh2s**) and "words, height and weight" (**whw2s**).

Shape to words. While our focus is on mapping from words to shape, it will be useful for several applications to do the reverse. To do so, we simply invert the linear regression in Eq. 2 of the w2s model to create a shape-to-words (**s2w**) model. Then given the 8 shape parameters of any body, we can predict a vector of word ratings.

3.3 Synthetic test data

To evaluate metric accuracy, we use the training meshes in a leave-one-out cross-validation approach. In addition to the ratings of each body (described above), we need the equivalent of "self-reported" height and weight to test the wh2s and whw2s models. To that end, we extract the ground truth height from the test meshes by taking the difference between the maximum and minimum vertex position in the vertical axis of the mesh. We calculated the weight of the training bodies by computing their volume and applying a standard approximation of body density. In the linear regression we use the cube root of weight as it is more linearly related to other measurements.

It is well known that people have systematic bias in self-reporting height and weight. For example, in one study, Spencer et al. [2002] found that men and women *overestimated* height by 1.23 (2.57) cm and 0.60 (2.68) cm, respectively (standard deviation in parentheses). Men and women also *underestimated* their weight by 1.85 (2.92) kg and 1.40 (2.45) kg, respectively, with heavier people underestimating more. With self reported measurements, one can correct for the known biases. For our test data, we assume the bias has been corrected and we add zero-mean Gaussian noise using the standard deviations above to simulate human self reporting error.

3.4 CAESAR test data

In addition to the synthetic bodies we randomly selected 50 female bodies from the CAESAR dataset [Robinette et al. 2002] and fit them with SMPL, using 300 principal components (effectively perfect reconstruction). The bodies were also fit with 8 components and the rendered images of these were rated as before but by different raters than those used in training the model. We use this dataset to test generalization performance and to evaluate how well our method captures realistic body shapes.

4 Model Evaluation

We evaluate our w2s model in terms of metric (geometric), measurement (anthropometric), and perceptual accuracy.

4.1 Metric evaluation (Reconstruction Error)

Metric analysis is performed on the training data using leave-one-out cross validation. For each gender, we train the w2s model 128 times, leaving out one body and its ratings each time. The ratings for the held-out bodies are used to predict the w2s body shape vector, giving us 256 predicted body shape vectors, each representing one of the synthetic bodies. Using SMPL we reconstruct the body meshes using each predicted body shape vector. We then compare the original synthetic bodies with the predicted bodies to quantify the prediction accuracy of our words-to-shape (w2s) model.

We define prediction accuracy in terms of "reconstruction error" (RE), which is the mean absolute distance between each vertex in the original body mesh and the corresponding vertex in the mesh that is reconstructed from the words. The RE is calculated for each of the 128 female and 128 male bodies. The results give an RE

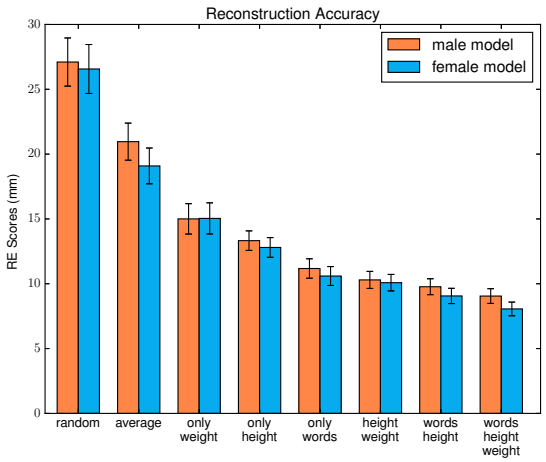

Figure 3: Reconstruction error. *Error in estimating synthetic female and male bodies using different methods of prediction: random, average, weight only, height only, words only, height+weight, words+height, words+height+weight. Whiskers correspond to the standard error (SE).*

of 10.595mm (SD = 8.233 mm) for female bodies and an RE of 11.011mm (SD=8.457) for male bodies.

We also evaluate linear models trained using various combinations of words, height and weight. Recall that height and weight here include realistic synthetic noise derived from [Spencer et al. 2002]. The results are summarized in Fig. 3. As a baseline we report accuracy using the average body as well as using body shapes randomly sampled from SMPL. We see that words alone can predict shape well even though the ratings contain no explicit metric information. Height and weight alone give reasonable metric accuracy, but combining them with words is even better. The words provide additional metric information.

The most accurate model is the whw2s model that uses words, height and weight. This gives an RE of 8.06 mm (SD=5.93) for female bodies and an RE of 9.05 mm (SD=6.42) for male bodies. For comparison, a commercial scanning solution using 10 Kinect frames has an error of 3.4 mm [Bogo et al. 2015] in a lab setting on different subjects.

To test the significance, we conduct dependent t-tests of the mean RE scores obtained from the cross validation procedure. The t-test reveals a significant difference between the mean RE obtained from the w2s model and the mean RE obtained from using words, height and weight together, $t = 6.686, p < 0.001$. Further, there is a significant difference between the RE obtained from the height and weight model and the RE scores obtained from the model with all three, $t = 5.545, p < 0.001$. The difference in RE between the w2s model and using only height and weight is not significant, $t = 0.986, p = 0.325$.

Given the similarity observed for RE in men and women (Fig. 3), in the remainder of the paper we report results only for women, unless otherwise stated.

Figure 4 shows examples of reconstructions of female body meshes from word ratings (column 2, "words"). The errors in the prediction are mostly at the extremities and can be attributed primarily to errors in estimating the height of the body. The qualitative shape is similar to the true body. Bodies predicted from just height and weight fail to capture body shape while the model combining

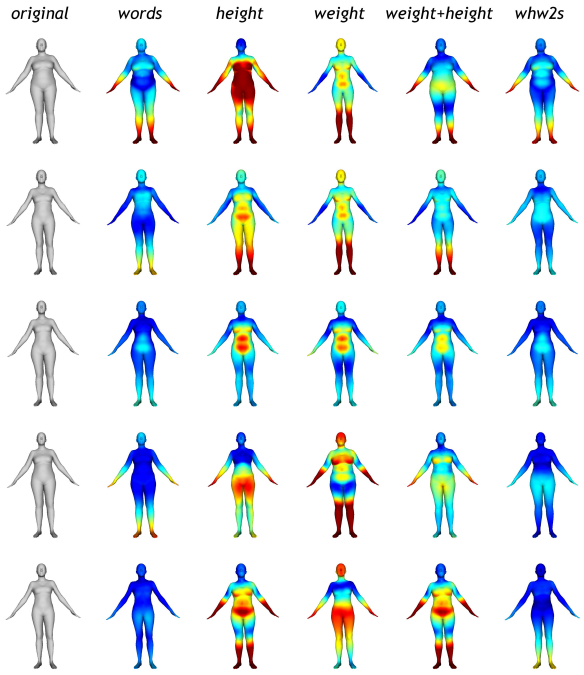

Figure 4: Words to shape estimation. *Selection of original and reconstructed female body meshes. The first column shows the synthetic body shape used to collect word ratings (rendered differently here). The color scale indicates the reconstruction error (RE); Blue = 0 mm and red > 30 mm. Column 2 shows the predicted mesh from words only. The meshes in columns 3-5 are predicted from height and weight without words. Column 6 shows the meshes predicted by combining words, height, and weight.*

words, height and eight (whw2s) results in predictions that are both visually similar to the truth and metrically more accurate.

Generalization. To evaluate the sensitivity of the whw2s model to a particular set of raters and bodies, we use the model trained with the synthetic dataset and test on the CAESAR test data (Sec. 3.4). The raters of the training and test data do not overlap. We evaluate the prediction of the 8-dimensional fit to the CAESAR data using the whw2s model. The RE for these bodies is 8.848mm, which is consistent with our results on held out synthetic data. This suggests that shape accuracy is not dependent on a particular set of raters and that the model generalizes to new raters and real bodies.

The first 8 PCs of the SMPL PCA space capture 96.56% of the variance in 2000 body shapes but are noticeably less detailed than bodies reconstructed using the full PCA space. For a real task, we want to reconstruct high-dimensional bodies accurately. To evaluate this we use the estimated bodies from the generalization experiment above and compute the RE to detailed realistic body shapes. To that end we use the 50 bodies from the CAESAR dataset reconstructed with 300 PCs, which represent 99.8% of the vertex variance.

The baseline RE between ground truth bodies represented with 8 and 300 PCs is 5.2mm. The RE between bodies reconstructed using whw2s with 8 PCs and the 300 PC ground truth is 10.653mm As expected, rating and estimating the 8-PCs shapes and comparing this to real body shape produces slightly higher errors (1.805mm) than the results above.

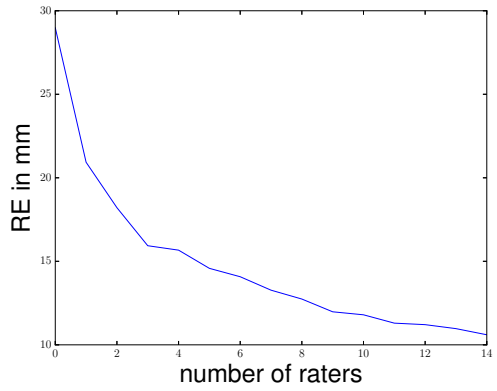

Figure 5: How many ratings are needed? *Reconstruction error (RE) in millimeters versus the number of ratings per image.*

full figured	long neck	high-breasted	slender
narrow hands	narrow feet	ample bosom	thin
desirably plump	short arms	big hands	angular
skinny legs	skinny arms	small-breasted	tall
pliantly slender	full bodied	slim waist	willowy
rounded stomach	high waist	big head	plump
cinched waist	flat chested	large breasts	
narrow shoulders	androgynous	full hips	

Table 2: Shape attributes extension. *New set of 30 shape attributes.*

How many ratings? For all experiments in this paper we use 15 ratings per word and per body. Figure 5 shows the effect of the number of ratings on accuracy; this uses all 30 words. Using only one rater does not produce metrically accurate results. It is the "wisdom of the crowd" that enables the accuracy. It is well known that averages of scalar ratings by the crowd can produce accurate metric judgements despite the fact that any individual rating is far from accurate [Treynor 1987]. We observe this behavior here. From the plot it appears that even more ratings could further reduce the error.

Model optimization. The relationship between body shape parameters and ratings is not always linear. We tried several non-linear regression methods and achieved slightly better results (lower RE) using support vector regression with an RBF kernel. The linear model, however, is presented here because it is nearly as good and is easy to interpret. We also used *ridge regression* for the w2s and obtained a reduction in RE to 9.77 mm for female bodies and 10.378 mm for male bodies. These results suggest that with more training data and more sophisticated models, lower errors may be possible.

Of our original 30 attributes, 27 were selected based on the findings of [Hill et al. 2015]. The meaning of some of the attributes in this set are similar, for example *heavy set* and *stocky*, which suggests that their ratings may be highly correlated. There are also many other ways to describe body shape. For the experiments below we collected 30 more words from literary texts and linguistic descriptions of female bodies (Table 2). We obtained ratings for these, with new raters, as before, resulting in a total of 60 words with 15 ratings for every training body.

We then search for a smaller set of words that compactly describes body shape. Using a greedy search algorithm and the w2s model with ridge regression, we rank words based on how much they improve the RE of held-out data. Figure 6 shows the best 30 words and how the RE decreases as each new word is added. Notice that already with the best 10 words the RE is 9.70 mm and with 25 it is 9.03 mm. Note that these errors are lower than using the original set of 30 words. As we see below, it is easy to add new words as needed for particular tasks.

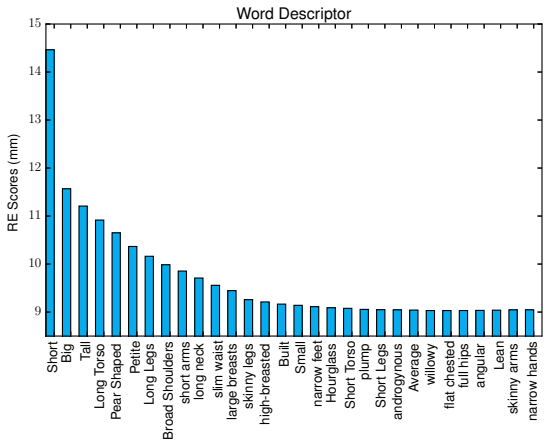

Figure 6: Words ranked by RE improvement. *Using greedy search, we ranked our extended set of shape attributes (60 words) to understand which words are most relevant for female body shape reconstruction. Here, words that start with a capital letter belong to the original set of shape attributes.*

4.2 Anthropomorphic Evaluation

For applications such as clothing sizing, measurements like lengths and girths are important. To evaluate errors in such anthropometric measurements, we extract the measurements from the reconstructed meshes by calculating distances between joint locations (e.g. upper arm length is calculated as the distance between the shoulder joint and the elbow joint) or by calculating circumferences around specific body parts such as hips, waist, or neck. Using the synthetic data and cross validation as above, we compute the average absolute errors between the anthropomorphic measurements of the predicted bodies and those of the true bodies.

Table 3 shows the mean absolute prediction error (MAE) for each measurement. Scalar ratings with words alone constrain measurements surprisingly well, reflecting the "wisdom of the crowd" phenomenon [Treynor 1987]. For example, the raters have no absolute cues about height, yet estimate it with an average error of 2.6 cm. This is equivalent to one standard deviation of self reported error in height [Spencer et al. 2002]. When noisy height and weight are added, the errors decrease significantly.

We also list the allowable error (AE) as specified by the US Army [Gordon et al. 1989]. AE is derived from the repeatability of expert human measurements of the body. Note that in Tsoli et al. [2014] it is clear that many sizing methods that use high-resolution scans have trouble achieving errors below the AE. While computed lengths and girths do not correspond directly to those in the Army study, these are provided in the table as a rough rule of thumb as to what would be a good error value for these measurements. We selected measurements from that study that are most similar to ours; there is not always a corresponding measure. In Tsoli et al. [2014] they report an error of 10.02 mm on 40 anthropometric measurements extracted from high-resolution CAESAR

	w2s		whw2s		
Measurement	MAE	SD	MAE	SD	AE
height	26.21	20.80	15.51	11.6	10
weight	4.21	3.19	1.87	1.42	
lower leg length	7.94	6.16	6.28	4.89	6
upper leg length	7.01	5.33	5.64	4.34	6
upper arm length	5.09	4.19	4.07	3.51	6
lower arm length	5.49	4.12	4.45	3.78	
neck length	2.86	2.18	2.52	1.92	
torso length	11.40	7.95	8.32	6.35	
shoulder width	6.67	4.75	4.21	3.41	
average linear	9.43	7.26	6.69	5.21	
calf girth	9.05	7.37	5.46	4.59	
thigh girth	17.50	15.00	12.17	9.76	6
waist girth	23.83	18.30	16.16	14.4	
chest girth	23.39	19.00	15.43	14.2	15
hip girth	25.34	20.2	14.44	11.9	12
neck girth	5.96	4.63	3.56	3.04	
head girth	8.26	6.48	5.32	3.77	
arm girth	9.77	7.63	5.48	4.49	
wrist girth	4.11	3.19	2.12	1.74	
average girth	14.14	11.34	8.90	7.57	

Table 3: Anthropomorphic measurement errors (women). *Mean absolute errors (MAE) and standard deviations (SD) for several body measurements: linear error in mm, weight in kg. A model using only words is already quite accurate. Errors go down when self-reported height and weight are used (here with simulated noise). AE refers to the "allowable error" (see text).*

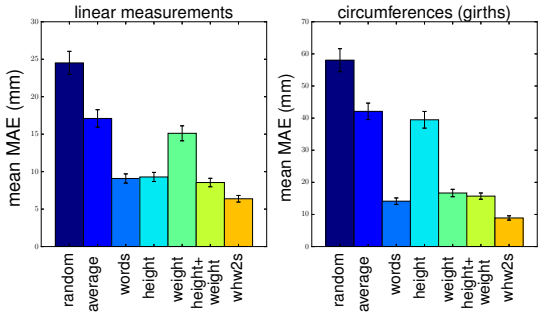

Figure 7: Anthropometric measurement errors. *This figure summarizes the measurement errors for linear measurements and circumferences (girths).*

Method	MSS	SD
Method 1 (scan fit)	5.265	1.403
Method 2 (w2s)	**5.537**	1.295
Average	3.659	1.761

Table 4: Perceptual study. *Mean similarity score (MSS) and standard deviation (SD) (7=extremely similar, 1= not similar at all) of different meshes compared to a photo of a person. The body shape created from words alone is visually as realistic as a body created from a high-resolution 3D scan of the person. Both are significantly more similar to the subject than an average body.*

scans. We do not have a full anthropometric measurement system, making a direct comparison impossible, but the average error on our subset of measurements is below 10 mm on synthetic bodies, suggesting that our accuracy may be sufficient to be useful.

Without the use of a scanner, and with noisy height and weight, whw2s estimates body shapes with errors close to the AE (and below in 2 of the 7 cases). If noiseless height and weight is known, then all errors drop below the AE with the exception of thigh girth. New attributes could be added to the model that focus on the thighs, likely reducing this error.

Figure 7 summarizes the results of the anthropometric analysis of different models. The two bar plots show the average measurement errors using the same models in figure 3.

4.3 Perceptual evaluation

In creating avatars, metric accuracy is not the only criterion for success. In fact it is easy to construct bodies that have low metric error but do not look like the subject of interest, and vice versa. Our perceptual evaluation tests the ability of the word-to-shape model to produce perceptually believable 3D digital bodies. In [Hill et al. 2015] they show strong agreement between ratings of photographs and ratings of 3D models constructed from the ratings of photographs. Here we take a different approach and test whether human subjects can tell the difference between bodies constructed from ratings of photos and those constructed from a high-resolution 3D scan.

We use two different methods for generating personalized digital bodies and compare the results in a similarity rating experiment. For Method 1 we scanned 6 human subjects with different body shapes using a high-resolution 3D scanner (3dMD, Atlanta, Georgia). Subjects gave informed, written, consent. We aligned a SMPL model to each of the scans by optimizing the pose and shape parameters to fit the scan data. For consistency with our w2s method, we optimized only the first 8 principal shape directions. We rendered images of the bodies as in the training data.

We also took a color digital photograph of each of the subjects. In Method 2 we had 30 MTurk users rate the photographs using the 30 words. We used w2s to estimate the body shape parameters from the ratings and generated the 3D body meshes. These were rendered as in Method 1. Figure 8 shows the data: the photograph and the two rendered bodies.

In a perceptual similarity study, 30 MTurk users rated the similarity between the photographs and 1) an average shape, 2) bodies from scans (Method 1), and 3) bodies from words (Method 2). Similarity was assessed using a 7-point Likert scale ranging from (1) not similar at all to (7) extremely similar. Raters rated a total of 18 similarity comparisons (6 models times 3 comparisons).

For each condition we compute the mean similarity score (MSS). The results are summarized in Table 4. Remarkably the w2s body is judged as slightly more similar to the image than a body fit directly to the 3D scan of the person, though the difference is not significant (paired t-test, t=1.297, p=0.251). Both methods produce bodies significantly more similar to the photograph than the average body (paired t-test).

This suggests that the visual shape ratings capture perceptually salient information about body shape. This could be important, for example, in clothing shopping where stylistic elements of clothing may be related to the perceived shape of the body in addition to measurements.

5 Applications

Visualizing word meaning. What does the word "pear-shaped" or "hourglass" mean in terms of 3D body shape? Our model allows us to visualize this, revealing what is "in our heads" when we

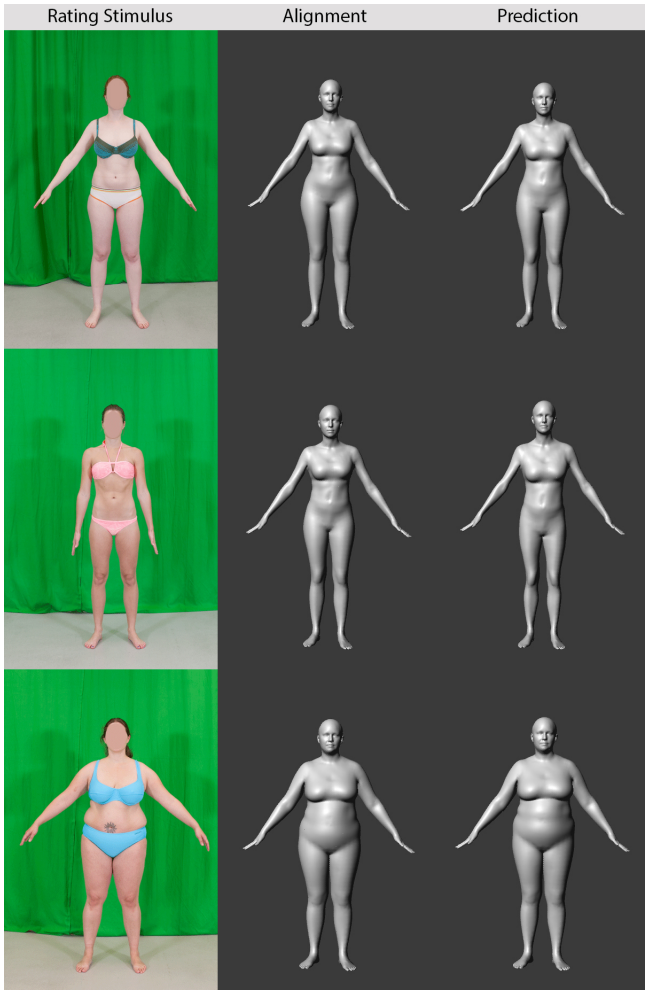

Figure 8: Perceptual Evaluation. *Example stimuli. (left) Photographs. (middle) Images of 3D meshes generated by aligning a SMPL model to high-resolution 3D body scans of the people in the photos (reposed to our standard pose). (right) Images of 3D body meshes predicted from the word ratings of the photographs using w2s. Mechanical Turk users rated the similarity between the photographs and the corresponding rendered meshes as well to an average body.*

use these words to describe bodies. Using the Gaussian model of ratings, we condition on a particular value of a rating as described in Sec. 3.2. Figures 1 and 9 show the most likely body shapes for which the rating of word displayed is set to 4 standard deviations from the mean of that rating. This exaggerates a particular dimension, revealing how raters interpret the word in terms of body shape. What is interesting is that the shapes are immediately recognizable as prototypical examples of the word.

Generation of bodies with attributes. We provide a web interface to allow people to create bodies using the attributes; see Fig. 10. Users can move attribute sliders to manipulate the body shape in real time, similar to previous systems [Jain et al. 2010]. One can use the sliders in a correlated way or can decouple them to manipulate particular attributes [Yumer et al. 2015; Jain et al. 2010]. Unconstrained attributes are estimated by conditioning the Gaussian model. The user can switch between editing with principal components or with attributes. When editing with PCs the user

Endomorph		
Soft body	Underdeveloped muscles	Round shaped
Mesomorph		
Hard	Muscular body	Mature appearance
Ectomorph		
Thin	Flat chest	Delicate build
Young	Lightly muscled	Stoop-shouldered
Tall		
Miss Wonderly		
Tall	*Pliantly slender*	*High-breasted*
Angular	*Narrow hands*	*Narrow feet*
Slender	**Long legs**	
Miles Archer		
Solidly built	Medium height	Wide shoulders
Thick neck	Jovial face	Heavy jawed

Table 5: Additional words. *We elicited additional ratings of the training bodies for these words associated with somatotypes and characters in the Maltese Falcon. Words in bold are already part of the original set of 30 words, while italics indicates words in the expansion set of 30 words (Table 2).*

sees the linguistic description of the body as a "word cloud," which visually illustrates how body shape and language are related. The website is available for research purposes at http://bodytalk.is.tue.mpg.de/.

Somatotypes. Bodies are described in many ways. Consider the classical breakdown into three body types: Mesomorph, Ectomorph, and Endomorph [Sheldon 1940]. Our goal is to uncover the 3D shapes that represent these body types. To do so, we took words associated with these shapes from [Som 2016] and collected new ratings of our 256 male and female training bodies. Note that we only used words and phrases associated with shape (Table 5); we did not use words associated with personality traits. Since the ratings are on the same bodies as before, we can simply expand our rating vectors to include the old and new words.

To construct the body shapes, we take the words associated with each type, set them to fixed values of 4.5 (out of 5), condition on these and estimate the remaining ratings. Figure 11 shows the reconstructed bodies. We believe this is the first time that realistic versions of these body shapes have been created from a statistical model of body shape.

3D Paparazzi. Can we take a photo of any person (e.g. a celebrity) and estimate a plausible 3D avatar? Here we take photos of famous people and submit them to MTurk for 15 ratings each using our original set of 30 words. We then reconstruct the body shape from the average ratings using the w2s model. Figure 12 shows a few of such images and the reconstructed body shape. We manually posed the body to be similar to the photo. The resemblance to the actual person is qualitatively reasonable when the person in the photo wears tight or minimal clothing. Clothing may obscure body shape making it harder for raters to judge (e.g. in Fig. 12 right, the person appears too slim). Results could be made more accurate using height and weight information readily available on the Internet.

Database search. Given a database of body shapes, our method is capable of indexing it with body descriptors and therefore, allowing descriptive queries over the bodies. The bodies in the CAESAR

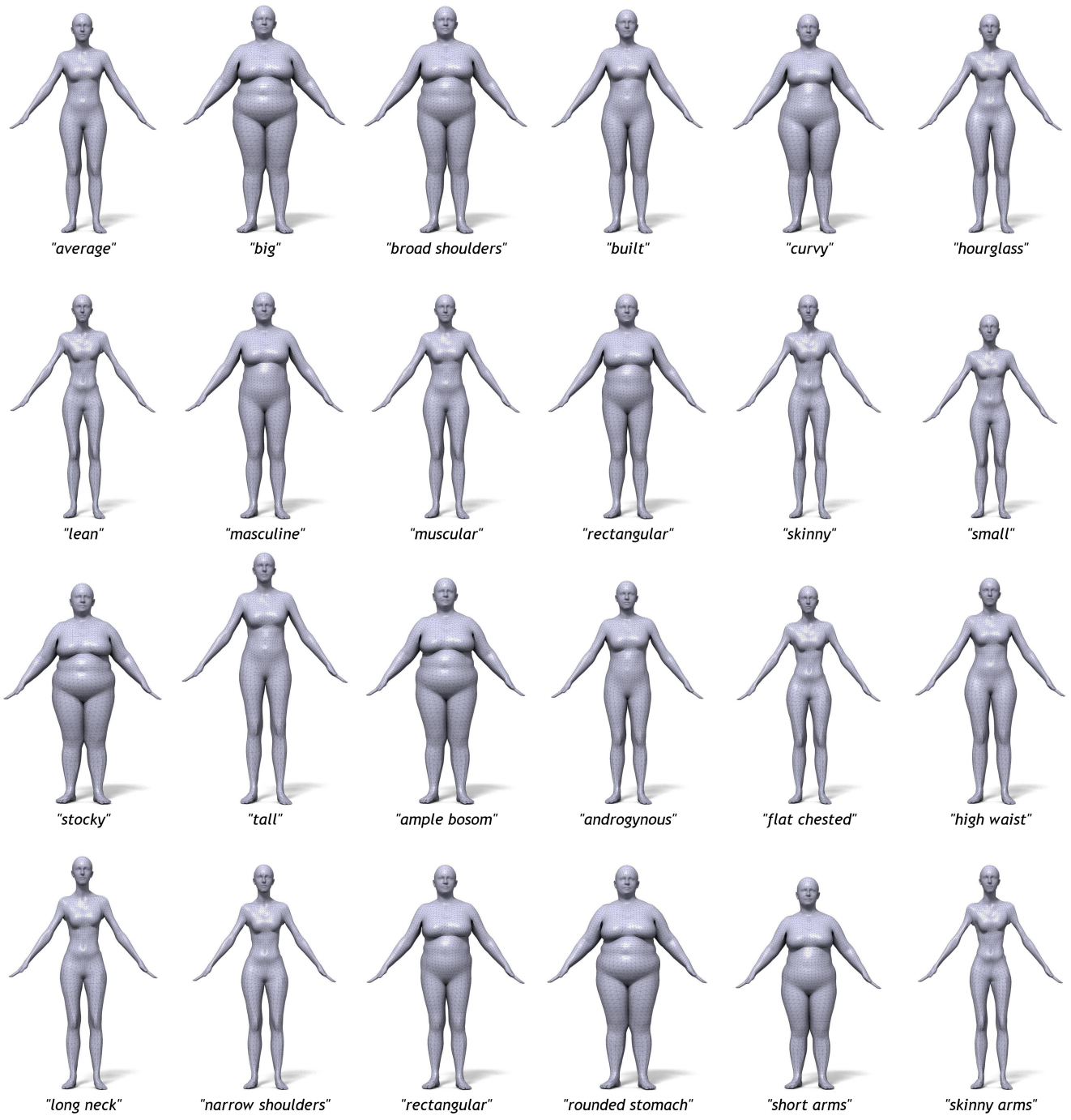

Figure 9: Visualizing words as shapes. *Prototype bodies are created by conditioning on an individual word rating and estimating the remaining ratings using correlations in the training ratings. We then generate the body shape using the rating vector with the w2s model (see text). These prototype bodies reveal the meaning that the crowd associated with each word.*

database are aligned using the SMPL model and, for each body, we generate shape attributes corresponding to our words. Unlike all the other experiments here, we do not do collect ratings for the CAESAR bodies. Instead we automatically generate ratings using the inverse shape-to-word (s2w) model, which takes body shape parameters and predicts the rating vector. We then store these words and their rating values in the database with each body.

Now it is possible to query the database in the usual ways. Figure 13 presents sample queries over the CAESAR database. Searches for a particular shape attribute return meshes of real bodies that all share this property but exhibit significant variation in other dimensions. Notice that the bodies associated with the search for "long legs" all have a similar slender body shape. This semantic search is quite different from searching on "inseam," which would have

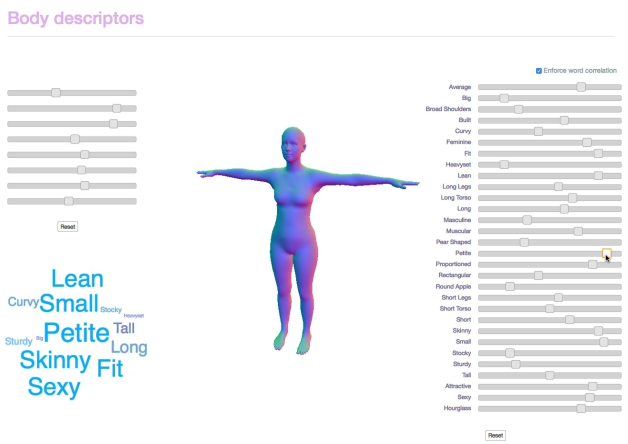

Figure 10: Web interface. *Screenshot showing the creation of a body (center) with attribute sliders (right) and principal component sliders (upper left). As the body shape changes so does the word cloud describing the body shape (lower left).*

returned bodies with wide range of body mass indexes (BMI). Instead, here we see the concept that the raters had of long legs. The "pear shaped" search returns bodies of varying height and BMI, that all have significant lower body fat as compared to their upper body. Finally, the search for "feminine" is maybe the most interesting. This is not a search that could be performed with standard measurements. It also reveals what our population of US MTurk users think is the feminine ideal; the women vary significantly in height and apparent ethnic background but have a slight build, small breasts, are slim, and are relatively fit. They are almost adolescent in appearance. This suggests that Body Talk, together with the CAESAR dataset, provides and interesting and powerful tool for exploring and understanding cultural ideals and attitudes towards body shape.

Bodies from books. We began the paper with a quote from the Maltese Falcon describing the character of Miles Archer. Here we take another, describing Miss Wonderly:

> *She was tall and pliantly slender, without angularity anywhere. Her body was erect and high-breasted, her legs long, her hands and feet narrow.*

We took the words from these two quotes (Table 5) and had the training set rated with them as above. Note that some of the words describe face shape, which may be correlated with body shape.

We added the words to the w2s model, set the ratings for these words to 3 standard deviations from the mean values, and estimated the most likely rating vector conditioned on these. Note that we rated "angular" but the text above reads "without angularity". In this case we set the rating to minus 3 standard deviations from the mean. The estimated body shapes are shown in Fig. 14 and both resemble the descriptions. Creating bodies from new words is as simple as having the training set rated with these words. There are several applications of this including creating avatars for games, printing physical models of characters for the blind, or casting actors with the physical characteristics described in a book or script.

6 Conclusions & Discussion

It is a longstanding question as to whether humans veridically represent the 3D world in the brain. It is unlikely that the brain would represent 3D shape either as a mesh as we do in the computer or

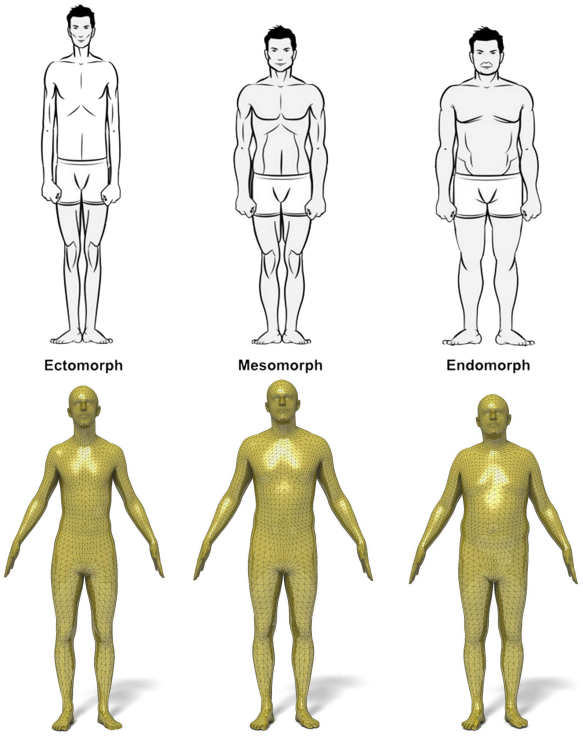

Figure 11: Somatotypes. *(top) Typical artist depiction of the somatotypes: Ectomorph, Mesomorph, and Endomorph. (Source: Wikipedia, Artist: Granito diaz, Creative Commons Attribution-Share Alike 4.0 International license). (bottom) Crowdshaping results from Body Talk.*

as a collection of words. Our results, however, suggest that humans do maintain a veridical 3D representation and can make the transformation from images to this representation and from the representation to a linguistic description. That is, our results suggest that brains maintain a general internal model of 3D shape that can mediate between perception and language. Moreover our results support the idea that this representation is metric. The results further suggest that the classical "wisdom of the crowd" extends to 3D body shape analysis. While any individual is a poor judge of shape (Fig. 5) the average of the crowd (here 15 raters) is accurate.

Limitations. Here we use a body shape space with 8 principal components. While this captures significant variance in the human population, some bodies may fall outside this space. Some words may correspond to rare shapes that are only captured by principal components with lower eigenvalues. Future work should explore extending the method to use more components.

We have assumed a linear relationship between ratings and shape coefficients. This is not true for all words. Some words like "skinny" are more categorical. Almost all bodies are "not skinny" and only bodies below some BMI are judged to be skinny. This suggests that a non-linear model would be better, but we suspect that more training data will be required to prevent overfitting.

Here we only show training bodies in a frontal view. This prevents rating some aspects of shape. Future work should present raters with side and/or rear views (either together or separately). This might result in a richer model of body shape. Future work should also evaluate whether pose influences shape perception (cf. [Sekunova et al. 2013]).

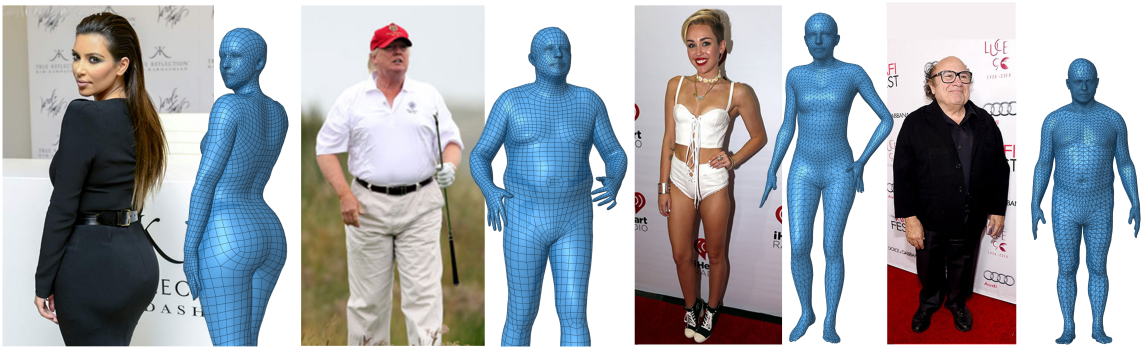

Figure 12: Celebrity bodies. *Crowdshaping a few famous bodies using photos from the Internet*[1]*. The rightmost example represents a "failure" case, where clothing obscures shape and the resulting body appears too slim.*

Our perceptual and metric evaluations focus on bodies with minimal clothing so that body shape is readily visible. How well can humans judge the shape of clothed people? The 3D paparazzi example suggests shape judgements of clothed people may be erroneous. This is not surprising since clothing is often worn to change the appearance of shape. Our tools could be used to study the effects of clothing on shape estimation, in particular how different clothing styles influence viewers' perceptions of shape.

Future work. Our results suggest the ability to predict anthropometric measurements from words. Future work should include a bigger study with more standard measurements. Since we have found that words contain metric information, we will explore whether it is possible to learn a direct mapping (possibly non-linear) from ratings to measurements, without first reconstructing a 3D body shape. For particular applications, like sizing jeans, additional words may be needed to capture the shapes of body parts (thighs, buttocks, etc.) that are relevant to fit.

Our web-based tool allows body creation using either word sliders or PCA sliders. We hypothesize that word sliders are easier to use for body creation by a single person since they are more perceptually intuitive than PCA sliders. We plan a user study to evaluate the speed and accuracy of creating bodies using different input methods.

We restricted our MTurk raters to English speakers in the United States. Clearly this study could be extended to other languages. More interesting, however, would be to use our tools to understand cross-cultural attitudes towards how bodies are perceived and how language is used to describe them.

Similarly, the way a body is rated may be correlated with the rater's own body and gender. In future work we will explore this by collecting body shapes of our raters. If true, collecting the body shape of the rater could be used to effectively normalize their ratings and increase prediction accuracy.

Our work here focuses exclusively on the body. There is previous work with faces that suggests the same methods will work there as well. More interesting, however, is the combination of faces and bodies. It is known that there are significant correlations between face and body shape but these are not well quantified. How much of body shape can be predicted by verbal descriptions of faces and vice versa is an open question.

[1]Credits: ©Marco Sagliocco / PR Photos; ©Ian MacNicol / Getty Images; ©Christopher Polk / Getty Images; ©Stefanie Keenan / Getty Images; respectively.

Conclusions. We have described a linear model that produces realistic 3D meshes of the human body from linguistic ratings of photographs. The approach uses ratings obtained by crowd sourcing and we show that the collection of such ratings, though not metric themselves, constrains body shape significantly. Our key observation is that these "crowdshaped" bodies are both perceptually and metrically accurate. Realistic 3D representations can be created using as few as 10 words. We demonstrate accuracy on the order necessary for many applications.

Body Talk can be used to generate a detailed 3D body shape if no scanning technology is available or applicable. For instance, the tool could allow users to generate a 3D representation of their own body by asking other users to rate a photograph of them. Our model allows us to visualize mental representations of human body shape and how this shape relates to our use of language. This work is useful for many fields (including psychology, medicine, cultural studies, art, etc.) where the study of body shape is important and simple tools are needed to create stimuli that probe human shape perception.

7 Acknowledgments

We thank Naureen Mahmood, Javier Romero, and Matt Loper for their advice in general and help with SMPL in particular.

References

ALLEN, B., CURLESS, B., AND POPOVIĆ, Z. 2003. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM transactions on graphics (TOG) 22*, 3, 587–594.

ALLEN, B., CURLESS, B., AND POPOVIĆ, Z. 2004. Exploring the space of human body shapes: Data-driven synthesis under anthropometric control. In *SAE International Proc. Digital Human Modeling for Design and Engineering Conference*.

ALLEN, B., CURLESS, B., POPOVIĆ, Z., AND HERTZMANN, A. 2006. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, SCA '06, 147–156.

ANGUELOV, D., SRINIVASAN, P., KOLLER, D., THRUN, S., RODGERS, J., AND DAVIS, J. 2005. SCAPE: Shape Comple-

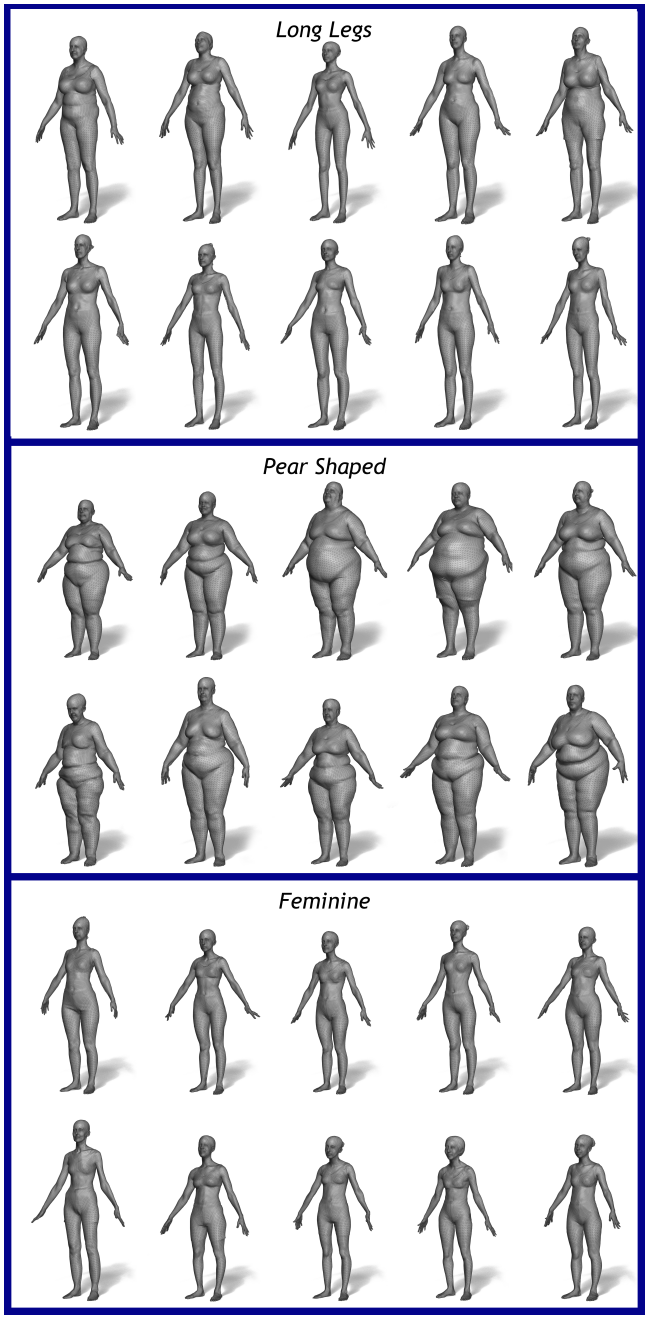

Figure 13: Database search with shape attributes. *Queries over the CAESAR dataset using "Long legs," "Pear shaped," and "Feminine." Displayed are the top semantic matches. Meshes correspond to the SMPL template mesh aligned to high-resolution CAESAR scans.*

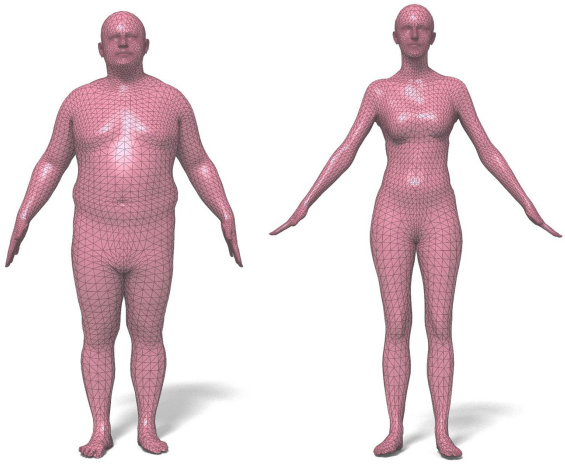

Figure 14: Bodies from books. *Using a character description in a book, we create the 3D avatar that matches their description. Here we create"Miles Archer" (left) and "Miss Wonderly" (right) from the Maltese Falcon.*

tion and Animation of PEople. *ACM Trans. Graph. (Proc. SIGGRAPH 24*, 3, 408–416.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, SIGGRAPH '99, 187–194.

BOGO, F., BLACK, M. J., LOPER, M., AND ROMERO, J. 2015. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Int. Conf. Comp. Vis. (ICCV)*, 2300–2308.

CARTER, J., AND HEATH, B. 1990. *Somatotyping: Development and applications*. Cambridge University Press, Cambridge.

CHAUDHURI, S., KALOGERAKIS, E., GIGUERE, S., AND FUNKHOUSER, T. 2013. Attribit: content creation with semantic attributes. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM, 193–202.

GORBER, S. C., TREMBLAY, M., MOHER, D., AND GORBER, B. 2007. A comparison of direct vs. self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Reviews 8*, 4, 307–326.

GORDON, C., CHURCHILL, T., CLAUSER, C., BRADTMILLER, B., AND MCCONVILLE, J. 1989. Anthropometric survey of US Army personnel: Methods and summary statistics 1988. Tech. rep., DTIC Document.

GREENACRE, M. 2007. *Correspondence analysis in practice (2nd ed.)*. Chapman and Hall/CRC.

GUY, S. J., KIM, S., LIN, M. C., AND MANOCHA, D. 2011. Simulating heterogeneous crowd behaviors using personality trait theory. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ACM, New York, NY, USA, SCA '11, 43–52.

HAMMETT, D. 1929. *The Maltese Falcon*. Alfred A. Knopf.

HASLER, N., STOLL, C., SUNKEL, M., ROSENHAHN, B., AND SEIDEL, H. 2009. A statistical model of human pose and body shape. *Computer Graphics Forum 28*, 2, 337–346.

HILL, M. Q., STREUBER, S., HAHN, C. A., BLACK, M. J., AND O'TOOLE, A. J. 2015. Exploring the relationship between body shapes and descriptions by linking similarity spaces. *J Vis. 15*, 12, 931.

HIRSHBERG, D., LOPER, M., RACHLIN, E., AND BLACK, M. 2012. Coregistration: Simultaneous alignment and modeling of

articulated 3D shape. In *European Conf. on Computer Vision (ECCV)*, Springer-Verlag, A. F. et al. (Eds.), Ed., LNCS 7577, Part IV, 242–255.

JAIN, A., THORMÄHLEN, T., SEIDEL, H.-P., AND THEOBALT, C. 2010. Moviereshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph. 29*, 6 (Dec.), 148:1–148:10.

KLARE, B., KLUM, S., KLONTZ, J., TABORSKY, E., AKGUL, T., AND JAIN, A. 2014. Suspect identification based on descriptive facial attributes. In *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, 1–8.

KLEINDESSNER, M., AND VON LUXBURG, U. 2015. Dimensionality estimation without distances. In *AISTATS*.

LI, H., VOUGA, E., GUDYM, A., LUO, L., BARRON, J. T., AND GUSEV, G. 2013. 3D self-portraits. *ACM Trans. Graph. (Proceedings SIGGRAPH Asia 2013) 32*, 6 (November).

LITTLE, A. C., ROBERTS, S. C., JONES, B. C., AND DEBRUINE, L. M. 2012. The perception of attractiveness and trustworthiness in male faces affects hypothetical voting decisions differently in wartime and peacetime scenarios. *Q J Exp Psychol 65*, 10 (May), 2018–2032.

LIU, T., HERTZMANN, A., LI, W., AND FUNKHOUSER, T. 2015. Style compatibility for 3D furniture models. *ACM Trans. Graph. (TOG) 34*, 4, 85.

LOPER, M., MAHMOOD, N., ROMERO, J., PONS-MOLL, G., AND BLACK, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 34*, 6 (Oct.), 248:1–248:16.

LUN, Z., KALOGERAKIS, E., AND SHEFFER, A. 2015. Elements of style: Learning perceptual shape style similarity. *ACM Trans. Graph. 34*, 4.

MATUSIK, W., PFISTER, H., BRAND, M., AND MCMILLAN, L. 2003. A data-driven reflectance model. *ACM Trans. Graph. 22*, 3 (July), 759–769.

O'DONOVAN, P., LĪBEKS, J., AGARWALA, A., AND HERTZMANN, A. 2014. Exploratory font selection using crowdsourced attributes. *ACM Trans. Graph. 33*, 4 (July), 92:1–92:9.

O'TOOLE, A., PRICE, T., VETTER, T., BARTLETT, J., AND BLANZ, V. 1999. 3D shape and 2D surface textures of human faces: The role of averages in attractiveness and age. *Image and Vision Computing 18*, 1, 9 – 19.

PARIKH, D., AND GRAUMAN, K. 2011. Relative attributes. In *Int. Conf. Comp. Vis. (ICCV)*, 503–510.

PONS-MOLL, G., FLEET, D., AND ROSENHAHN, B. 2014. Posebits for monocular human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2345–2352.

REID, D., AND NIXON, M. 2013. Human identification using facial comparative descriptions. In *Biometrics (ICB), 2013 International Conference on*, 1–7.

REID, D. A., NIXON, M. S., AND STEVENAGE, S. V. 2014. Soft biometrics; human identification using comparative descriptions. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 36* (June), 1216–1228.

ROBINETTE, K., BLACKWELL, S., DAANEN, H., BOEHMER, M., FLEMING, S., BRILL, T., HOEFERLIN, D., AND BURNSIDES, D. 2002. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Tech. Rep. AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory.

SEKUNOVA, A., BLACK, M., PARKINSON, L., AND BARTON, J. J. S. 2013. Viewpoint and pose in body-form adaptation. *Perception 42*, 2, 176–186.

SEO, H.-W., AND MAGNENAT-THALMANN, N. 2003. An automatic modeling of human bodies from sizing parameters. In *In Proceedings of the 2003 Symposium on Interactive 3D Graphics*, ACM Press, 1926.

SEO, H.-W., CORDIER, F., AND MANGNENAT-THALMANN, N. 2003. Synthesizing body models with parameterized shape modifications. In *In Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, Eurographics Association, 120125.

SHAPIRO, A., FENG, A., WANG, R., LI, H., BOLAS, M., MEDIONI, G., AND SUMA, E. 2014. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*.

SHELDON, W. H. 1940. *The Varieties of Human Physique (An Introduction to Constitutional Psychology)*. Harper and Brothers.

SIGAL, L., MAHLER, M., DIAZ, S., MCINTOSH, K., CARTER, E., RICHARDS, T., AND HODGINS, J. 2015. A perceptual control space for garment simulation. *ACM Trans. Graph. 34*, 4 (July), 117:1–117:10.

2015. SMPL model. http://smpl.is.tue.mpg.de/. Accessed: 2015-12-10.

2016. Somatotypes. http://users.rider.edu/~suler/somato.html. Accessed: 2016-01-14.

SPENCER, E. A., APPLEBY, P. N., DAVEY, G. K., AND KEY, T. J. 2002. Validity of self-reported height and weight in 4808 epicoxford participants. *Public Health Nutrition 5* (8), 561–565.

SUMNER, R. W. 2005. *Mesh modification using deformation gradients*. PhD thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science.

TALTON, J. O., GIBSON, D., YANG, L., HANRAHAN, P., AND KOLTUN, V. 2009. Exploratory modeling with collaborative design spaces. *ACM Trans. Graph. 28*, 5, 167.

TREYNOR, J. L. 1987. Market efficiency and the bean jar experiment. *Financial Analysts Journal 43*, 3 (May/June), 50–53.

TSOLI, A., LOPER, M., AND BLACK, M. J. 2014. Model-based anthropometry: Predicting measurements from 3D human scans in multiple poses. In *Proceedings Winter Conference on Applications of Computer Vision*, IEEE, 83–90.

WEISS, A., HIRSHBERG, D., AND BLACK, M. 2011. Home 3D body scans from noisy image and range data. In *Int. Conf. Comp. Vis. (ICCV)*, IEEE, Barcelona, 1951–1958.

XIA, S., WANG, C., CHAI, J., AND HODGINS, J. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Trans. Graph. (TOG) 34*, 4, 119.

YUMER, M. E., CHAUDHURI, S., HODGINS, J. K., AND KARA, L. B. 2015. Semantic shape editing using deformation handles. *ACM Trans. Graph. (Proceedings of SIGGRAPH 2015) 34*.

ZHOU, S., FU, H., LIU, L., COHEN-OR, D., AND HAN, X. 2010. Parametric reshaping of human bodies in images. *ACM Trans. Graph. 29*, 4 (July), 126:1–126:10.