HEIGHT AND WEIGHT ESTIMATION FROM UNCONSTRAINED IMAGES

Can Yilmaz Altinigne

School of Computer and Comm. Sciences (IC)
École Polytechnique Fédérale de Lausanne
1015 Switzerland

Dorina Thanou, Radhakrishna Achanta

Swiss Data Science Center (SDSC)
École Polytechnique Fédérale de Lausanne
1015 Switzerland

ABSTRACT

We address the difficult problem of estimating the weight and height of individuals from pictures taken in completely unconstrained settings. We present a deep learning scheme that relies on simultaneous prediction of human silhouettes and skeletal joints as strong regularizers that improve the prediction of attributes such as height and weight. Apart from imparting robustness to the prediction of attributes, our regularization also allows for better visual interpretability of the attribute prediction. For height estimation, our method shows lower mean average error compared to the state of the art despite using a simpler approach. For weight estimation, which has hardly been addressed in the literature, we set a new benchmark.

***Index Terms*—** Biometrics, deep learning, height and weight prediction, skeletal joint prediction, segmentation, interpretability, regularization.

1. INTRODUCTION

Height and weight estimation of individuals from images is a difficult problem, recurrent in many application domains such as surveillance, pedestrian traffic study for urban planning, automated garment fitting in online stores, and autonomous driving. The problem becomes even more challenging when images are taken in unconstrained settings, namely, in arbitrary lighting conditions, with uncalibrated cameras, with unknown distance from the subject, with unknown distance of the camera from the ground, and without a reference object of known dimensions.

In this work, we study the problem of weight and height estimation, in totally unconstrained environments, from single, frontal, full-body images. We build on state-of-the-art deep learning techniques, in particular the well-known U-Net architecture [1], and extend it to predict the height and the weight of an individual. In order to regularize the network, we force it to simultaneously predict the human silhouette (hereafter, the *mask*), as well as the skeletal joint locations (hereafter, *joints*). Such an additional requirement ensures that the network does not converge to simple statistics such as the mean value. Instead, it learns meaningful, silent features that are implicitly correlated to the height and the weight of the individual. In order to achieve this, besides the losses related to the regression of the attribute(s) such as height or (and) weight, we introduce two additional losses that take into consideration the reconstruction of the masks and the joint locations. We learn the parameters of our network in an end-to-end fashion. Experimental results on the IMDB dataset and Reddit images confirm that our proposed method outperforms state-of-the-art algorithms for height and weight estimation.

In summary, we introduce a novel method that predicts attributes like height and weight of individuals from single and unconstrained images without using any prior information related to characteristics of the individual such as age, gender, or face features. Our method avoids converging to trivial solutions such as the mean due to the additional regularizing prediction terms for mask and joint prediction. Our method allows us to interpret the quality of the attribute prediction since it is visually correlated to the prediction of the mask and joints. Finally, as an added contribution, to the best of our knowledge, with this paper we set the reference baseline for weight prediction using a publicly available database.

2. PRIOR WORK

There are several methods in the literature that deal with the prediction of different physical attributes of human such as height, weight, and body-mass index (BMI) using image data from uncalibrated cameras. BenAbdelkader and Yacoob [2] deploy a linear model that relies on manually labeled keypoint locations in the image, for height prediction. Dey et al. [3] estimate height differences of people in each image, and create a height difference graph from a photo collection. They predict the height by exploiting a prior distribution on this graph.

In addition to face and body images, anthropometric measurements such as body width and body area have also been used for both height and weight estimation. Linear combination of several anthropometric measurements are used for human weight estimation in [4, 5]. Also, various bone lengths are used for height estimation with linear regression models in [6, 7]. Moreover, Rativa et al. [8] apply typical regression methods such as SVM, Gaussian Process and Neural

Networks to several anthropometric measurements in order to predict height and weight of a person. More recently, Jiang and Guo [9] use frontal body images with labeled height and weight value to predict BMI. They first extract several anthropometric measurements by detecting joints and body contours which is similar to our research except the part of measurement extraction, and they use these measurements as inputs to regress BMI using Support Vector and Gaussian Process Regression models. Unlike their two-step approach of estimating height and weight, which is more error-prone, we use a single end-to-end approach that predicts attribute values at the same time as joint and masks in a single step, leading to a robust predictions despite the unconstrained nature of the input images.

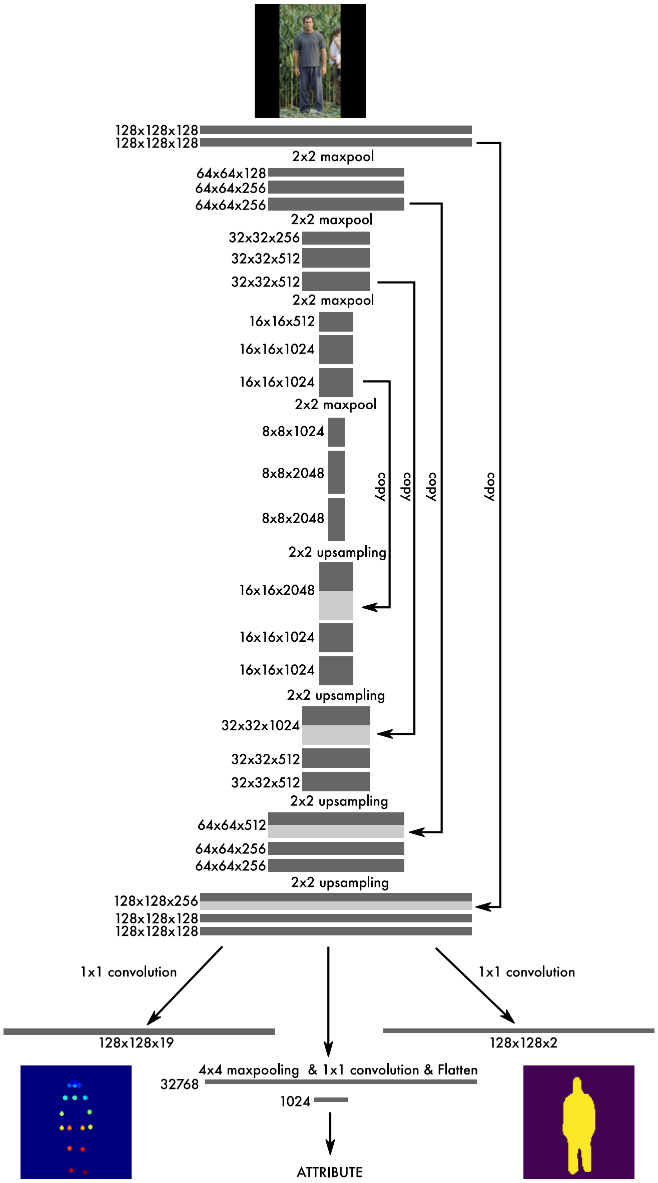

Fig. 1. *Figure explaining our network architecture in detail. As can be seen, there are three heads, one for predicting attribute(s) like height and weight, and two others for human silhouette segmentation and skeletal joint detection.*

Finally, deep learning in which we are mainly interested in in this work, has been recently applied for height or weight estimation. Gunel et al. [10] train a ResNet model [11] combined with a Light CNN structure [12] for height estimation. However, their training set consists of many multi-person, non-full body face and body images. Bieler et al. [13] present nearly two times lower error than Gunel et al. [10] on the same images using gravity as a reference to predict height values. However, they use a static camera to record a sequence of images (video) and they make strong assumptions that permit their model to predict height when free fall motion (jumping, running etc.) is recorded. In a different line of work, Dantcheva et al. [14] train a ResNet-50 network structure [11] with only face images as input to predict height, weight and body mass index of people. Nguyen et al. [15] use 400 uncalibrated full body images to predict human weight using Support Vector Regression and Ridge Regression models. However, they require Kinect-based [16] depth information as input along with the full body images.

3. OUR APPROACH

As stated previously, we predict the human contour mask as well as skeletal joints when regressing heights and weights. In this section we explain the proposed architecture, the datasets we use, as well as how we train the network.

3.1. Model Structure

The main structure is borrowed from the U-Net model of Ronnenberger et al. [1], with the exception that we use a higher number of feature channels in the encoding part (and correspondingly in the decoding part) of the network. This is depicted in Fig. 1. The input to the model is an RGB image with size 128×128. Our network has three output heads, one each for silhouette segmentation, skeletal joint prediction, and attribute prediction. Depending on the need or the training data, the attribute head can predict either height, or weight, or both. The predictions generated by the mask and the joint heads regularize the prediction of the attribute head, leading to more accurate results.

The segmentation head outputs masks of size $128 \times 128 \times 2$. The number of channels in this output corresponds to background and foreground pixels. The joint prediction head has dimensions of $128 \times 128 \times 19$. Each of the first 18 channels is responsible for predicting the location of one of the 18 joints, while the last one is responsible for non-joint locations. The attribute head takes the last layer of U-Net part as input to an adaptive max pooling layer to accommodate varying image sizes. Then, 1×1 convolution is performed to obtain a

$32 \times 32 \times 32$ output. This output is fed to a feed-forward network with one hidden layer with 1024 neurons and a dropout layer to obtain prediction of attribute(s).

3.2. Dataset Preparation

We use frontal full-body images in the IMDB dataset used by Gunel et al. [10] for height estimation. The dataset has nearly 109,000 images with height values for the salient person in each image. However, this dataset includes a very large number of images with multiple people and non-full body footage. We therefore remove these from the training, test and validation sets with the help of a face detection library [17]. In the end, our training, test, and validation set consist of 31,000, 6200, and 2000 randomly split images respectively.

Due to the lack of a complete, publicly available dataset for weight estimation, we crawl Reddit [18] for images and accompanying weight data. We extract in total 4400 images with weight data and perform similar data cleaning as for the height estimation images. From this, we create a training set of 3000 images, a test set of 900 images, and a validation set has 500 images, for the task of weight estimation.

We resize each image in the training data such that the longer edge is of 128 pixels. We pad the image symmetrically along the shorter edge to make the width equal to the height. To prepare the groundtruth for skeletal joint prediction, we use the joint locations predicted by OpenPose [19] with a small Gaussian spread, to avoid a single pixel associated with each joint. To prepare the groundtruth for the human segmentation mask, we use the output of Mask R-CNN [20]. Since the groundtruth is generated automatically, it is to be noted that the groundtruth may have errors introduced by these pretrained models. However, since the aim is to predict height and weight, using skeletal joint and segmentation mask predictions as regularizers, the errors in this groundtruth can be tolerated.

3.3. Training

We train the network with three losses, one corresponding to each output head, namely, for predicting skeletal joints $\hat{J}$, segmentation mask $\hat{M}$, and attributes of $\hat{a}$ corresponding to height or weight (or age) as the case may be. We use the cross-entropy loss $\mathcal{L}_J$ for joint prediction, i.e.,

$$\mathcal{L}_J = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\mathbf{1}_{J_i \in C_c} \log\left(p[\hat{J}_i \in C_c]\right), \qquad (1)$$

where n is the total number of training images, $C = 19$ is the number of channels corresponding to possible joint and non-join locations, and C_c is the class label of channel c. $\mathbf{1}_{J_i \in C_c}$ is a binary indicator that is 1 only if J_i is the correct class label C_c for the joint location c of image i, $\hat{J}_i$ denotes the channelwise $argmax$ value in the joint prediction output (128x128x19), and $p[\hat{J}_i \in C_c]$ is the predicted probability that $\hat{J}_i$ is of class C_c.

We use the dice-loss $\mathcal{L}_M$ for the mask prediction, which is defined as

$$\mathcal{L}_M = \frac{1}{n}\sum_{i=1}^{n} 1 - \frac{2 \times \hat{M}_i \circ M_i}{\hat{M}_i + M_i}, \qquad (2)$$

where M_i and $\hat{M}_i$ are respectively the true and predicted mask of image i, and $\circ$ denotes the element-wise multiplication. The height and weight prediction are quantified by the mean absolute error (MAE) $\mathcal{L}_A$, i.e.,

$$\mathcal{L}_A = \frac{1}{n}\sum_{i=1}^{n} |\hat{a}_i - a_i|, \qquad (3)$$

where $\hat{a}_i, a_i$ are respectively the predicted and the true value of the corresponding attribute of image i.

Finally, the total loss for height and/or weight estimation are defined as the sum of the above losses, i.e.,

$$\mathcal{L} = \mathcal{L}_M + \mathcal{L}_J + \mathcal{L}_A. \qquad (4)$$

While, the three losses can have a different relative weight, in our case, we weigh them equally.

4. EXPERIMENTS AND RESULTS

A problem we face is the dearth of datasets that provide both height and weight groundtruth for individuals. Since our reference datasets, described in Sec. 3.2, have only height or weight information but not both, we train the network once each to predict these attributes separately.

We train the height estimation network with only losses related to mask, joint, and height prediction on the IMDB dataset with 31000 images. For both height and weight estimation case, we use the *Adam* [21] optimizer with a learning rate of $1e-4$. We set the batch size to 16, train the network for 50 epochs and retain the model checkpoint with the best validation loss for testing.

Similarly, we train the weight estimation network with losses related to mask, joint, and weight prediction. Since both models use the same U-Net structure, we initialize the weights in U-Net part with the weights of the best model for height estimation. We give as an input the training set of the weight dataset which consists of 3000 images. We again set the batch size to 16, train the network for 100 epochs using *Adam* [21] optimizer with a learning rate of $1e-4$, and retain the model checkpoint with the best validation loss for testing.

Our results for height and weight estimation are shown in Table 1. We note that we obtain 6.13 cm mean absolute error on our test set, which includes 6200 full body single person images from the IMDb dataset. We compare our results with Gunel at al. [10] which is the only work that uses unconstrained RGB images directly, without any additional

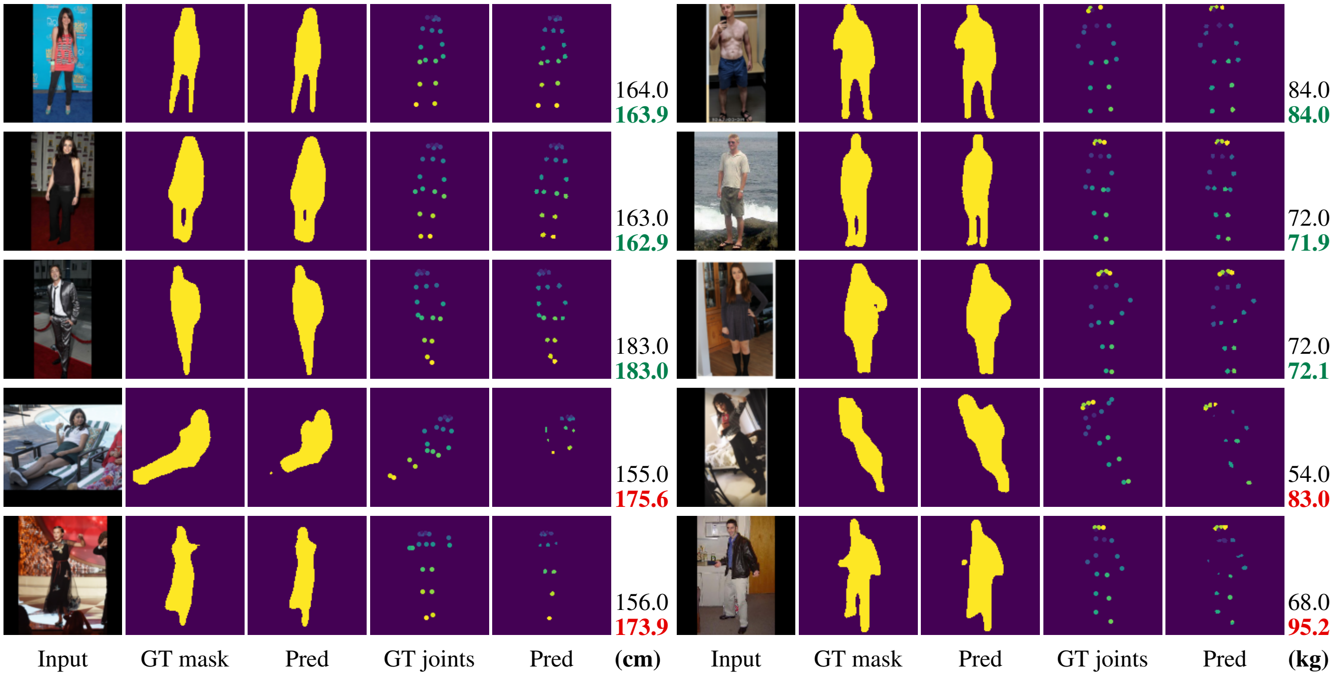

Fig. 2. *Examples of visual interpretability of height and weight estimation. The numbers in* **bold** *are attribute values predicted using our approach, with those in* ***green*** *being much closer to the groundtruth (GT) as compared to the* ***red*** *ones. As can be observed in the lower two rows of images, when height and weight prediction show a large error, the corresponding mask and joint predictions tend to be poor.*

assumptions. Gunel et al. [10] use face images in addition to full body images. When using only fully body, only face, and both body and face images, they report an MAE of 6.40, 6.25, and 6.06, respectively. As a baseline, we use the MAE with respect to the mean height of the images in the training set. For the sake of a fair comparison, we train the model of [10] on our cleaner full body dataset using the best hyperparameters that they present in their paper from scratch. In this setting, we perform better than Gunel et al. [10] despite using a simpler approach, less computation, and less training data (see Table 1).

Height model	MAE	Weight model	MAE
Baseline (Mean)	8.06	Baseline (Mean)	15.60
Gunel et al. [10]	6.67		
Ours	**6.13**	**Ours**	**9.80**

Table 1. *Comparison of MAE of height estimation (in centimeters) and that of weight estimation (in kilograms).*

As shown in Table 1, our weight estimation method has 9.80 kg mean absolute error on the test set of Reddit dataset consisting of 900 images, which is much lower in comparison to a prediction based on the mean weight of the test data. The closest approach to our weight estimation is due to Jiang and Guo [9], who estimate BMI (rather than height and weight). Unfortunately, the authors were not able to provide us their code or dataset, so we were unable to provide a direct comparison with their method. To the best of our knowledge, we are the first to perform weight estimation from unconstrained images using deep learning.

The results of Table 1 prove that our approach of predicting height and weight using joint and mask prediction as regularizers is a reliable method. Apart from regularization, the prediction of masks and joints also lends some degree of interpretability to the results, which is due to the close correlation of the quality of height and weight prediction with that of the mask and joints. As can be observed in Fig. 2, when the attribute prediction is good, the mask or joint prediction tend to be accurate, and vice versa.

5. CONCLUSION

We present a new scheme for the estimation of attributes like height and weight from completely unconstrained images. In our method, we make use of skeletal joint detection and human contour-based segmentation tasks as regularizers. Not only does this improve the results of attribute estimation, as can be seen from our comparisons against the state of the art, it also brings visual interpretability to the predictions. For height estimation we outperform the state of the art, while for weight estimation, which has rarely been addressed in the literature, we set a new benchmark.

6. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[2] Chiraz BenAbdelkader and Yaser Yacoob, "Statistical body height estimation from a single image," in *8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1–7.

[3] Ratan Dey, Madhurya Nangia, Keith W Ross, and Yong Liu, "Estimating heights from photo collections: A data-driven approach," in *2nd ACM Conference on Online Social Networks*, 2014, pp. 227–238.

[4] Carmelo Velardo and Jean-Luc Dugelay, "Weight estimation from visual body appearance," in *4th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2010, pp. 1–6.

[5] Donald Adjeroh, Deng Cao, Marco Piccirilli, and Arun Ross, "Predictability and correlation in human metrology," in *2010 IEEE international workshop on information forensics and security*. IEEE, 2010, pp. 1–6.

[6] John Albanese, Andrew Tuck, José Gomes, and Hugo FV Cardoso, "An alternative approach for estimating stature from long bones that is not population- or group-specific," *Forensic science international*, vol. 259, pp. 59–68, 2016.

[7] Izzet Duyar and Can Pelin, "Body height estimation based on tibia length in different stature groups," *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, vol. 122, no. 1, pp. 23–27, 2003.

[8] Diego Rativa, Bruno JT Fernandes, and Alexandre Roque, "Height and weight estimation from anthropometric measurements using machine learning regressions," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–9, 2018.

[9] Min Jiang and Guodong Guo, "Body weight analysis from human body images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2676 – 2688, 2019.

[10] Semih Günel, Helge Rhodin, and Pascal Fua, "What face and body shapes can tell about height," *arXiv:1805.10355*, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[13] Didier Bieler, Semih Günel, Pascal Fua, and Helge Rhodin, "Gravity as a reference for estimating a person's height from video," *arXiv:1909.02211*, 2019.

[14] Antitza Dantcheva, Francois Bremond, and Piotr Bilinski, "Show me your face and I will tell you your height, weight and body mass index," in *24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3555–3560.

[15] Tam V Nguyen, Jiashi Feng, and Shuicheng Yan, "Seeing human weight from a single rgb-d image," *Journal of Computer Science and Technology*, vol. 29, no. 5, pp. 777–784, 2014.

[16] Zhengyou Zhang, "Microsoft kinect sensor and its effect," *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[17] "OpenCV face detection," `https://github.com/opencv/opencv/tree/master/samples/dnn`.

[18] "Reddit Progresspics - Show us your body transformations," `https://www.reddit.com/r/progresspics/`.

[19] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *arXiv:1812.08008*, 2018.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.