What Face and Body Shapes Can Tell Us About Height

Semih Günel[1], Helge Rhodin[1,2], Pascal Fua[1]
[1]Computer Vision Lab, EPFL, Lausanne, Switzerland
[2]UBC, Vancouver, Canada
{semih.gunel, pascal.fua}@epfl.ch, rhodin@cs.ubc.ca

Abstract

Recovering a person's height from a single image is important for virtual garment fitting, autonomous driving and surveillance. However, it is also very challenging without absolute scale information. Here, we examine the rarely addressed case, where camera parameters and scene geometry are all unknown. Under this circumstances, scale is inherently ambiguous, and height can only be inferred from those statistics that are intrinsic to human anatomy and can be estimated from images directly, such as articulated pose, bone-length proportions, and facial features. Our contribution is twofold. First, we create a new human-height dataset that is three magnitudes larger than existing ones, by mining explicit height labels and propagating them to additional images through face recognition and assignment consistency. Second, we test a wide range of machine learning models (linear, shallow, and deep models) to capture the relation between image content and human height. We also show that performance is predominantly limited by dataset size. Our central finding is that height can only be estimated with large uncertainty. The remaining high variance demonstrates that the geometrically motivated scale ambiguity persists into the age of deep learning, which has important implications for how to pose monocular reconstruction, such as 3D human pose estimation, in a scale invariant way.

1. Introduction

Estimating people's height from a single image is needed in areas such as subject identification for surveillance purposes, pedestrian distance estimation for autonomous driving, and automated garment fitting in online stores. However, since people's apparent height is affected by camera distance and focal length, assessing someone's real height only from the image is difficult. Therefore, the current 3D human pose estimation methods output only normalized pose, leaving the absolute scale of the person undetermined. [33, 34, 35, 36]. Scale can be recovered given a prior knowledge of the scene, for instance, if the location of the ground plane is given [1, 2, 3, 4] or where objects of known height are visible [3, 5]. Often, however, the scene content is unknown and can not be modified. In this paper, we investigate whether correlations in human appearance, shape and body proportions alone suffice for estimating absolute height. Our approach is inspired by medical studies that suggest correlations between body proportions and human height [6, 7, 8, 9, 10, 11, 12, 13, 14, 15], such as the ratio of the tibia length to the whole body or the head to shoulders ratio, but it remains unclear if these can be estimated from images. Existing algorithms can only operate under very specific conditions. For example, comparisons to the average height have been used in [16] to infer people's sizes from group pictures. The method of [17] is the only one we know of that can operate on single uncalibrated RGB images and without prior knowledge. However, it relies on manually supplied keypoints and does not generalize well to real images.

Recent deep learning successes suggest that any well-defined problem can be addressed given enough data. To examine this, we build a large dataset and study how much human pose and facial features can tell us about height. We observe that the posterior height estimate improves on the baselines, deeper networks outperform shallow architectures, accuracy increases with the database size and that both face and pose are viable cues. Closer investigations reveal that facial features provide the most important cues, which hints at determining gender and ethnicity being dominant factors.

Our contribution is empirical in nature with fundamental implications for the theoretical design of future height and pose estimation approaches. First, we have confirmed that, in this area as in many others, Deep Nets can be trained end-to-end and exceed the state-of-the-art in accuracy [17]. However, this requires a training set several orders of magnitude larger than those used in previous studies, which are featuring only a handful of subjects [18, 19, 20]. We have therefore introduced a novel and practical approach to mining a large training dataset via label propagation and we will

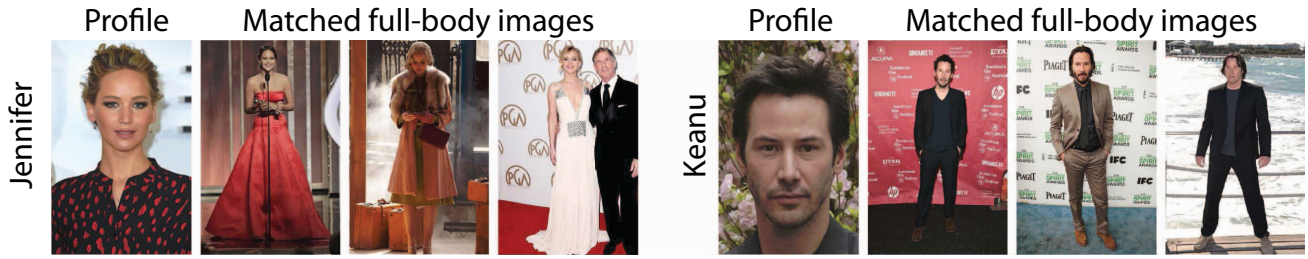

Figure 1. **Examples from *IMDB-100K*.** Profile images have been matched to additional images. Thereby, height labels on portrait images are propagated to all the assigned images.

make the resulting database publicly available.

Second, and most importantly, we have demonstrated that the geometric reasoning about the difficulty of the monocular height estimation problem remains valid in the age of deep learning. Despite the deep network with large capacity, a big dataset, and well-defined input-label pairs the estimates remain uncertain with a mean absolute error of 5.56cm, which only slightly exceeds the predictive power of the population mean, attaining 5.91cm.

2. Related Work

There are several algorithms that can infer age [21, 22, 23, 24] or emotional state [25, 26] from single images with high reliability, often exceeding that of humans, in part because these are not affected noticeably by scale ambiguities. By contrast, there are far fewer approaches to estimating human size and we review them briefly here.

Geometric height estimation. The height of standing people can be estimated geometrically from a single image under some fairly mild assumptions. This can be done by finding head position and foot contact through triangulation when the camera height and orientation in relation to the ground plane is known [1, 2, 4], computing the vanishing point of the ground plane and the height of a reference object in the scene [3, 27], or accounting for the height of multiple nearby reference objects [3, 5]. However, the necessary knowledge about camera pose, ground plane, and feet contact points is often unavailable.

Height from camera geometry. Without external scale information, object size is ambiguous according to the basic pinhole camera model. In practice, lenses have a limited depth of field, which shape-from-defocus techniques exploit [28, 29] to estimate distance. It can be used to guess depth orderings in a single image. However, a focal sweep across multiple images or a specialized camera [30] is required for metric scale reconstruction.

Height from image features. In [16], face position and size are used to measure relative heights in pixels first in group pictures and then in image collections featuring groups. Absolute height is estimated from the network of relative heights by enforcing consistency with the average human height, which is effective but only for group photos. Closest to our approach are the data-driven ones of [17] and [31]. The former uses a linear-regressor to predict height from keypoint locations in the input image. The results of an anthropometric survey [32] are used to train the regressor. However, even though the keypoints are supplied manually, the results on real images barely exceed what can be done by predicting an average height for all subjects. By contrast, our DeepNet regressor is non-linear, can learn a much more complex mapping that accounts for the uncertainty of image-feature extraction, does not require manual annotation of keypoints, and yields better results. [31] use a deep network similar to ours, but test facial features as input only and train on a much smaller dataset, leaving open the question whether deep nets can succeed given sufficiently large datasets and facial as well as full-body information.

Our network architecture is inspired by deep networks used for 3D human pose prediction [33, 34, 35, 36, 37, 38, 39, 40]. However, we will show that training on the existing 3D pose datasets with a handful of subjects is insufficient, which was our incentive for creating a larger one.

Height from body measurements. Medical studies suggest that the height of an individual can be approximated given ratios of limb proportions [6], absolute tibia length [15], foot length [7], and the ratio of head to shoulders [8, 9]. Also human perception of height seems influenced by head to shoulders ratio, which suggests a real link between head to shoulders ratio to actual height [10, 11, 12]. There is also a body of anthropological research about inferring the living height of the individual from the length of several bones in their skeletons, which indicates that height can be approximated given the size of some body parts [13, 14, 15].While these studies indicate that height estimation should be possible from facial and full-body measurement, there is no easy way to obtain them from single uncalibrated images and it is not known how naturally occurring feature extraction error influences accuracy. In particular, the often mentioned absolute length measurements cannot be inferred directly from 2D images.

Figure 2. **Identity matching.** Samples from the processed *IMDB-100K* dataset, with an overlay of the assigned subjects' 2D pose, head detection, identity and height annotation. In favor of reliable assignments opposed to false assignments, some persons remain unassigned.

3. Method

Our goal is to estimate human height, $h \in \mathbb{R}$, from a single RGB image, $I \in \mathbb{R}^{3\times n\times m}$, without prior knowledge of camera geometry, viewpoint position, or ground plane location. This setting rules out any direct measurement and requires statistical analysis of body proportions and appearance from the images only. We therefore follow a data-driven approach and infer the relationship between image content and human height through machine learning.

To make the method independent of scene-specific content, we first localize people in the image and then learn a mapping $f_\theta(\bar{I})$ from image crops $\bar{I}$ that tightly contain the target subject. To this end, we first introduce a diverse dataset of cropped image-height pairs, $D = \{(\bar{I}_i, h_i)\}_i^N$, with N examples (Sec. 3.1). Then, we explore different image features and neural network architectures to infer parameters θ of $f_\theta(I)$ that robustly predict height h given a new input $\bar{I}$ (Sec. 3.2).

3.1. Dataset Mining

Although existing 3D pose datasets contain human height information, they are limited to a handful of subjects [18, 19, 20] (Human3.6Million, HumanEva and MPII-INF-3DHP). On the other hand, datasets build from web content and comprising anonymous individuals [41, 42, 43] (MPII-2D, BBC-Pose, COCO) do not include height information. We therefore create the first large scale human height dataset containing more than a handful of subjects. Our dataset includes 12.104 subjects in the final version with known height. We started from a medium-sized one containing people of known height, which we then enlarged using face-based re-identification and pruned by enforcing label consistency and filtering on 2D pose estimates. Fig. 1 depicts the result.

Initialization. We used the IMDB website as our starting point. As of February 2018, it covers 8.7 million show-business personalities.[1] To find heights of people and corresponding images, we crawled the most popular 100.000 actors.[2] We found 12,104 individuals with both height information and a profile image involving a single face.

Augmentation. IMDB also has more than a million images taken at award ceremonies and stills from movies, including full-body images of our 12,104 individuals. Although there are associated labels specifying the actors present in the image, these labels do not specify location of the person in the image, which makes the association of height labels to a single person in image potentially ambiguous, especially if there are several people present.

Formally, let I be an image that should be labeled and S_I be the subject labels given by IMDB. We ϕ run a face detection algorithm [44], which returns a set K_I of detected individuals and for each $k \in K_I$ the head location in terms of a bounding box and a feature vector $\mathbf{v}_k$ that describes the appearance compactly. When there is only one person in the image, we can directly attribute the associated height information to the detected subject. This has enabled us to create a first annotated dataset of 23,024 examples, which we will refer to as *IMDB-23K*.

The same strategy was used to create the IMDB-WIKI dataset to learn age from facial crops [21]. However, we also need the rest of the body and want to create a richer database by also using images with several people and multiple detections. To associate labels and detections in such cases, we compute from the profile image of each subject $j \in S_I$ a facial descriptor $vs._l$ and store its euclidean distance from comparable descriptors for all detections $\{\mathbf{v}_k\}_{k\in K_I}$ in image I. This yields a distance matrix $\mathbf{D} \in \mathbb{R}^{n_K\times n_S}$ between all the $\mathbf{v}_k$ and $vs._j$ descriptors. To match one $\mathbf{v}_k$ to a specific $vs._j$, we make sure that $\mathbf{D}[k, j]$ is smaller than all other distances in row k and column j, that is $\mathbf{v}_k$ is the closest match to all $vs.$ in the list of subjects and, similarly, $vs._j$ is the closest match to all $\mathbf{v}$ in the list of detections. In practice, we apply an additional ratio test to ensure the assignment is reliable: We assign $vs._j$ to the best matching feature vector $k^* = \text{argmin}_k \mathbf{D}[k, j]$, but only if the quotient $q = \frac{\mathbf{D}[k^*,j]}{\min_{k\neq k^*}\mathbf{D}[k,j]}$ is smaller than

[1] https://www.imdb.com/pressroom/stats/

[2] http://www.imdb.com/search/name?gender=male,female

Figure 3. **Qualitative evaluation.** Results on test set of *IMDB-test* shown as prediction/ground-truth in centimeters.

$\tau = 0.9$. That means best match must be significantly better than the second-best match to be accepted. This produces a much larger set of 274,964 image-person pairs with known height, which we will refer to as *IMDB-275K*. We show a few examples in Fig. 2.

To estimate the accuracy of our assignments in *IMDB-275K* we randomly select 120 images which include multiple faces, where assigning and identity to faces is non-trivial and possibly erroneous. Out of the 331 IMDB labels in 120 images, we assigned 237 labels to faces inside the images, where only 5 of the assignments were wrong. We also repeated the same experiment for *IMDB-23K*, where assignment is much easier. We again selected 120 random images and check the accuracy of the assignments. We observed only a single mismatch. Overall this corresponds to an estimated label precision of 98.0% and recall of 70.1%.

Filtering and preprocessing. As discussed in the Related Work section, previous studies suggest that full-body pose and bone-length relations contain scale information. Therefore, we run a multi-person 2D pose estimation algorithm [45] on each dataset image I and assign the detected joints to the subject whose estimated head location as predicted by the face detector [44] is closest to the one estimated by the 2D pose extraction algorithm. The 2D joints are then used to compute image crops $\bar{I}$ that tightly enclose the body and head. The face is similarly cropped to $\tilde{I}$, as shown in the left side of Fig. 4.

Finally, we automatically exclude from *IMDB-275K* images missing upper body joints or whose crop is less than 32 pixels tall. This leaves us with 101,664 examples, which we will refer to as the *IMDB-100K* dataset. We also applied this process to *IMDB-23K*. In both cases, we store for each person the annotated height h, a face crop $\tilde{I}$, a facial feature vector $\mathbf{v}$, a pose crop $\bar{I}$, a set of 2D joint locations $\mathbf{p}$, and the gender if available.

Splitting the dataset. We split *IMDB-100K* into three sets, roughly in size 80k, 15k and 5k images for training, testing and validation, respectively.

3.2. Height Regression

Since there is little prior work on estimating human height directly from image features, it is unclear which features are the most effective. We therefore tested a wide range of them. To the face and body crops, $\tilde{I}$ and $\bar{I}$, discussed in Section 3.1, which we padded to be 256×256, we added the corresponding 2D body poses, in the form of 2D locations of keypoints centered around their mean and whitened, along with 4096-dimensional facial features computed from the last hidden layer of the VGG-16-based face recognition network of [46].

Given all these features, we tested the three different approaches to regression depicted by Fig. 4. A baseline that resembles the state-of-the art:

- *Linear*. Linear regression from the pre-computed 2D pose and facial features vectors, as in [17].

As well as two, more complex, neural network architectures:

- *ShallowNet*. Regression using a 4-layer fully connected network as used in [35]. *ShallowNet* operates on the same features as *Linear*.
- *DeepNet*. Regression using a deeper and more complex network to combine fine-grained facial features with large-scale information about overall body pose and shape. It uses two separate channels to compute face and full body features directly from the body and face crops, respectively, and uses two fully connected layers to fuse the results, as depicted in Fig. 4. By contrast to **ShallowNet**, we train this network end-to-end and thereby optimize the facial and full-body feature extraction networks for the task of human height estimation using MSE Loss. To allow for a fair comparison, we use the same VGG architecture in the face stream[46]. For the full body one, we utilize a ResNet [47].

4. Evaluation

We now quantify the accuracy brought about by our estimation and try to tease out the influence of its individual

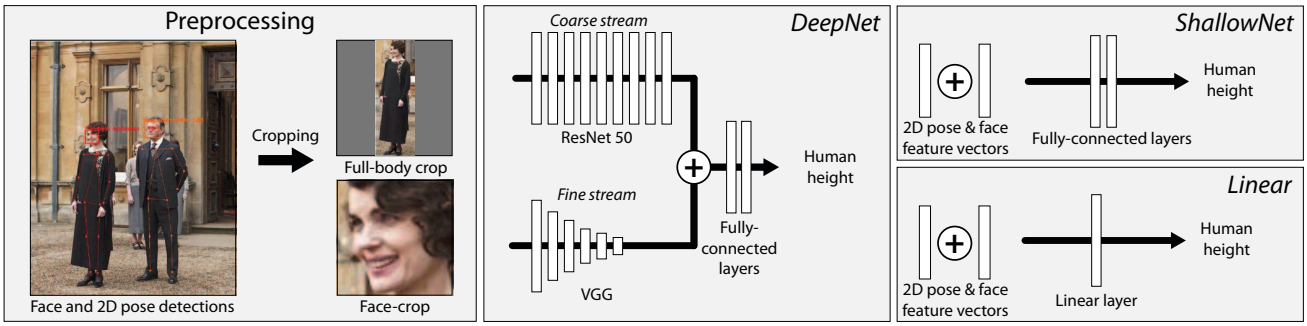

Figure 4. **Deep two-stream architecture.** Humans are automatically detected and cropped in the preprocessing state. We experiment with *DeepNet*, a two-scale deep convolutional network that is trained end-to-end, and a simple *ShallowNet*, that operates on generic image features, and with *Linear*, which is a simple Linear Regression.

components, dataset mining, and network design. We also show some example results on Fig. 3 for the most popular actors from the test split of *IMDB-100K*.

Metrics. We report height estimation accuracy in terms of the mean absolute error (MAE) compared to the annotated height in cm. We also supply cumulative error histograms.

Independent test set. To demonstrate that our training dataset is generic and that our models generalize well, we created *Lab-test*, an in-house dataset containing photos of various subjects whose height is known precisely. Since it was acquired completely independently from IMDB, we can be sure that our results are not contaminated by overfitting to specific poses, appearance, illumination, or angle consistency. *Lab-test* depicts 14 different individuals with 10 photos each. Each one contains a full body shot in different settings, sometimes with small occlusions to reflect the complexities of the real world. The subjects are walking in different directions, standing, or sitting. Individuals span diverse ethnicities from several European and Asian countries, and heights ranging from 1.57 to 1.93 in meters.

Baselines. We compare *DeepNet* against the following baselines in order of increasing sophistication:

- *ConstantMean*. The simplest we can do, which is to directly predict the average height of *IMDB-100K*, which is 170.1 centimeters.
- *GenderMean*. Since men are taller than women on average, gender is a predictor of height. We use the ground-truth annotation as an oracle and the gender-specific mean height as the prediction, which are 166 cm for woman and 180 cm for man.
- *GenderPred*. Instead of using a gender oracle, we train a network whose architecture is similar to *DeepNet* to predict gender instead of height and again use the gender-specific mean height as the prediction.
- *Linear* and *ShallowNet* as introduced in Sec. 3.2.
- *PoseNet*. We re-implemented the method of [20] that predicts 3D human pose in absolute metric coordinates after training on the Human3.6M dataset [18]. Height information is extracted from the predicted bone lengths from head to ankle. To accommodate for the distance from ankle to the ground, we find the optimal constant offset between the predicted height and the ground truth height on *IMDB-100K*, in the least squares sense.

4.1. Comparative Results

We report our mean accuracies on *IMDB-100K* and *Lab-test* along with those of the baselines at the top Tab. 1(a). *DeepNet*, which is our complete approach, outperforms them on both, with *GenderPred* being a close second. This shows that knowing the gender is indeed a strong height predictor, but it is not the only one. To confirm this, we retrain *DeepNet* for men and women separately and compared

	IMDB-100K			*Lab-test*
Method	all	women	men	all
ConstantMean	8.25	7.46	9.22	11.0
GenderPred	6.61	6.28	7.12	9.26
PoseNet [20]	-	-	-	10.65
DeepNet (ours)	**6.14**	**5.88**	**6.40**	**9.13**
GenderMean	5.91	5.63	6.23	8.66
DeepNet (gender-specific)	**5.56**	**5.23**	**6.03**	**8.53**

(a)

	Regression type		
Input features	Linear	*ShallowNet*	*DeepNet*
Body crop only	7.56 / 11.10	7.10 / 10.40	6.40 / 9.43
Face crop only	6.49 / 10.25	6.31 / 9.99	6.25 / **8.87**
Body and Face	6.40 / 10.2	6.29 / 9.92	**6.14** / 9.13

(b)

Table 1. **Mean Absolute Error (MAE) in cm on *IMDB-100K* and *Lab-test*.** (a) Comparison against our baselines. (b) Ablation study, accuracies are given in *IMDB-100K* / *Lab-test* format.

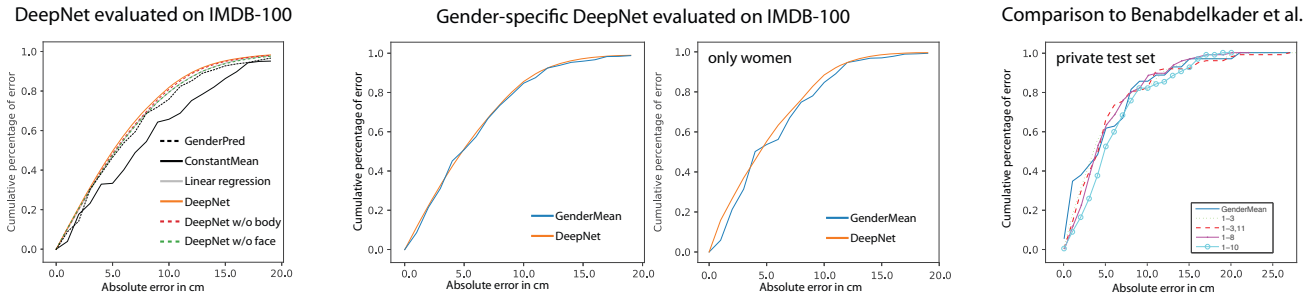

Figure 5. **Cumulative error analysis.** *DeepNet* improves significantly and consistently across all error segments on *GenderPred* and other baselines (first plot). Compared to *GenderMean*, the improvement is small but consistent, except for the 4cm mark (second plot). The gender specific analysis reveals that improvements on *GenderMean* are more pronounced for women (third plot compared to the second plot). The rightmost (fourth) plot gives an indirect comparison to Benabdelkader et al. [17], who evaluated on a different private dataset but used the same metric and baseline.

its accuracy to that of *GenderMean*, shown at the bottom of the table. Our full approach improves upon gender specific means, somewhat unexpectedly, more for women than for men.

The accumulated error histograms in Fig. 5 show that these findings are consistent. *DeepNet* improves significantly on *GenderPred* (first plot), consistently on *GenderMean* across different error ranges (second plot, 4cm mark is the only exception), and that the performance on women is systematically better. Furthermore, the rightmost plot of Fig. 5 allows an indirect comparison to [17]. The authors of [17] evaluate on a different private test set. However, none of their variants exceeds the *GenderMean* baseline consistently, while ours does, particularly for women.

The second baseline we discuss is *PoseNet*, which does not do particularly well, presumably because it has not learned the vast variety of possible body shapes because it has been trained on many images but all from only five subjects. We demonstrate in a subsequent experiment that orders of magnitudes more subjects are needed.

In Table 1(b), we report the results of an ablation study in which we ran the three versions of our algorithm—*Linear*, *ShallowNet*, and *DeepNet* introduced in Section 3.2—on the full dataset, on the faces only, or on the body only. In all cases, *DeepNet* does better than the others, which further indicates that it also outperforms the state-of-the-art algorithm [17], which *Linear* emulates.

Most conclusions drawn from experiments on *IMDB-100K* are confirmed on *Lab-test*. Surprisingly, using both body and faces helps on *IMDB-100K*, but not on *Lab-test* where using the faces only is best. We suspect that the poses in *Lab-test* are more varied than in *IMDB-100K* and, therefore, face features generalize better. Furthermore, there is also a wider spread of heights in *Lab-test* and other biases due to its smaller size, which might contribute to this behavior. When either pose or facial features are used, facial features are superior on both datasets, which further suggest that facial features provide the most important cues. This hints at determining gender and ethnicity being dominant factors, but also head-size could play a role.

Overall, the seemingly strong predictive power of gender and relatively small improvements brought by full-body and facial features demonstrates that monocular scale estimation remains a largely ill-posed problem.

4.2. Dataset Size and Quality

In Fig. 6, we plot the accuracy of our model as a function of the size of the training set. It clearly takes more than 10,000 to 20,000 images to outperform *GenderMean*. Interestingly, it seems to take more images for men than women, possibly due to the larger variance in men height. This indicates the results we report here for men might not be optimal yet and would benefit from using an even larger training set.

The method of [31] has been trained on a much smaller set 1400), and reports a much larger error (7.7 cm MAE). Albeit errors are reported on a different test set, this confirms the finding that much larger datasets are needed to train deep nets on the height estimation problem.

To estimate the accuracy of our assignments in *IMDB-275K* we randomly select 120 images which include multiple faces, where assigning identity to faces is non-trivial and possibly erroneous. Out of the 331 IMDB labels in 120 images, we assigned 237 labels to faces inside the images, where only 5 of the assignments were wrong. We also repeated the same experiment for *IMDB-23K*, where assignment is much easier. We again selected 120 random images and check the accuracy of the assignments. We observed only a single mismatch. Overall this corresponds to an estimated label precision of 98.0% and recall of 70.1%.

5. Discussion and Limitations

We have made the best possible effort of creating a large enough dataset, validated that the proposed label assignment is effective introducing negligible label noise (dot vs.

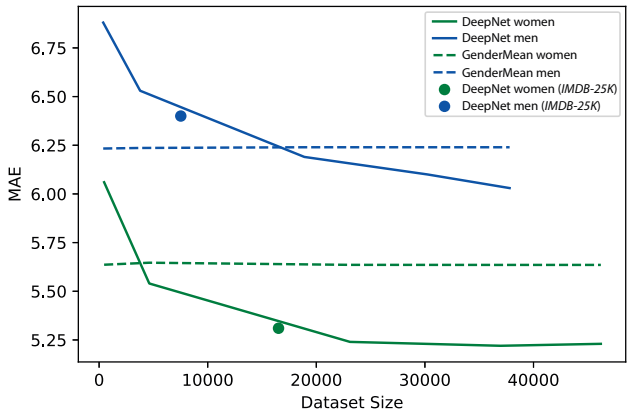

Figure 6. **Accuracy as a function of the training dataset size.** We plot separate curves for men and women.

line in Fig. 6), and ensured that all available information (face, full-body, and articulated pose) is accessible to the network. In spite of using a deep network with large capacity and a big dataset of well-defined input-label pairs to train it, there remain substantial uncertainty in our size estimates. This also reflects the fact that scale ambiguity remains a difficult computer vision—and even human vision—problem and that additional context cues remain needed. One future direction of research will be making sure the dataset is consistent, possibly by validating the annotations in group pictures [16]. Furthermore, if some annotations can be identified as unreliable, this could be modeled by incorporating a confidence value during training and prediction.

6. Conclusion

With 274,964 images and 12,104 actors, the dataset we created is the largest one to date for height estimation. The label association it provides can be used not only for height estimation but also to explore other properties of human appearance and shape. We experimented with different network architectures that improve on current height estimation algorithms. However, some scale ambiguity remains and is unlikely to be solved by machine learning alone.

Our findings have several implications for future work in the area of height prediction. As there remains a substantial amount of height uncertainty, human 3D pose estimation algorithms should not be evaluated in metric space, as is often done, but after scale normalization; the inevitable inaccuracies in height estimation should be evaluated separately. Furthermore, if absolute height is desired, a large dataset must be used for training purposes to cover the large variations in human shape, pose and appearance. Finally, it is important to use facial features on top of full-body information for height regression.

References

[1] Y. Guan, "Unsupervised human height estimation from a single image," *Journal of Biomedical Science and Engineering*, 2009.

[2] X. Zhou, P. Jiang, X. Zhang, B. Zhang, and F. Wang, "The Measurement of Human Height Based on Coordinate Transformation," in *ICIC*, 2016.

[3] J. Vester, "Estimating the Height of an Unknown Object in a 2D Image," Master's thesis, KTH, 2012.

[4] S. Li, V. Nguyen, M. Ma, C. Jin, T. Do, and H. Kim, "A simplified nonlinear regression method for human height estimation in video surveillance," in *EURASIP Journal on Image and Video Processing*, 2011.

[5] J. Ljungberg and J. Sönnerstam, "Estimation of human height from surveillance camera footage -a reliability study," Master's thesis, KTH, 2008.

[6] D. Adjeroh, D. Cao, M. Piccirilli, and A. Ross, "Predictability and correlation in human metrology," in *IEEE International Workshop on Information Forensics and Security*, 2010.

[7] A. Zaslan, M. Yaar, I. Can, I. Zaslan, H. Tugcu, and S. Koç, "Estimation of stature from body parts," *Forensic Science International*, 2003.

[8] T. Shiang, "A statistical approach to data analysis and 3-D geometric description of the human head and face." in *Proceedings of the National Science Council, Republic of China. Part B, Life sciences*, 1999.

[9] K. Kato and A. Higashiyama, "Estimation of height for persons in pictures." in *Perception & psychophysics*, 1998.

[10] D. Re., L. Debruine, B. Jones, and D. Perrett, "Facial Cues to Perceived Height Influence Leadership Choices in Simulated War and Peace Contexts," *Evolutionary Psychology*, 2013.

[11] G. Mather, "Head and Body Ratio as a Visual Cue for Stature in People and Sculptural Art," *Perception*, 2010.

[12] C. Burton and N. Rule, "Judgments of Height from Faces are Informed by Dominance and Facial Maturity," *Social Cognition*, 2013.

[13] R. Wilson, N. Herrmann, and L. Jantz, "Evaluation of Stature Estimation from the Database for Forensic Anthropology," *Journal of Forensic Sciences*, 2010.

[14] J. Albanese, A. Tuck, J. Gomes, and H. Cardoso, "An alternative approach for estimating stature from long bones that is not population- or group-specific," *Forensic Science International*, 2016.

[15] I. Duyar and C. Pelin, "Body height estimation based on tibia length in different stature groups," *Am J Phys Anthropo*, 2003.

[16] R. Dey, M. Nangia, W. Ross, W. Keith, and Y. Liu, "Estimating Heights from Photo Collections: A Data-Driven Approach," in *ACM conference on Online social network*, 2014.

[17] C. BenAbdelkader and Y. Yacoob, "Statistical Body Height Estimation from a Single Image," in *Automated Face and Gesture Recognition*, 2008, pp. 1–7.

[18] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[19] L. Sigal, A. Balan, and M. J. Black, "Humaneva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *International Journal of Computer Vision*, 2010.

[20] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision," in *International Conference on 3D Vision*, 2017.

[21] R. Rothe, R. Timofte, and L. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," in *International Journal of Computer Vision*, 2016.

[22] R. Malli, M. Aygun, and H. Ekenel, "Apparent Age Estimation Using Ensemble of Deep Learning Models," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[23] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," *Neurocomputing*, 2016.

[24] X. Wang, R. Guo, and C. C. Kambhamettu, "Deeply-Learned Feature for Age Estimation," in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015.

[25] D. Dagar, A. Hudait, H. Tripathy, and M. Das, "Automatic emotion detection model from facial expression," in *2016 International Conference on Advanced Communication Control and Computing Technologies*, 2016.

[26] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. How individual face parts contribute to successful emotion recognition," *PLOS One*, 2017.

[27] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[28] G. Mather, "Image blur as a pictorial depth cue," in *Proc. R. Soc. Lond. B*, 1996.

[29] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[30] T. Georgiev, , Z. Yu, A. Lumsdaine, and S. Goma, "Lytro Camera Technology: Theory, Algorithms, Performance Analysis," in *Multimedia Content and Mobile Devices*, 2013.

[31] A. Dantcheva, F. Bremond, and P. Bilinski, "Show me your face and I will tell you your height, weight and body mass index," 2018.

[32] C. Gordon, T. C. land C.E. Clauser, B. Bradtmiller, and J. McConville, "Anthropometric survey of us army personnel: methods and summary statistics 1988," Anthropology Research Project Inc Yellow Springs OH, Tech. Rep., 1989.

[33] D. Tome, C. Russell, and L. Agapito, "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image," in *arXiv preprint, arXiv:1701.00295*, 2017.

[34] A.-I. Popa, M. Zanfir, and C. Sminchisescu, "Deep Multitask Architecture for Integrated 2D and 3D Human Sensing," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[35] J. Martinez, R. Hossain, J. Romero, and J. Little, "A Simple Yet Effective Baseline for 3D Human Pose Estimation," in *International Conference on Computer Vision*, 2017.

[36] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-Time 3D Human Pose Estimation with a Single RGB Camera," in *ACM SIGGRAPH*, 2017.

[37] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-Net: Localization-Classification-Regression for Human Pose," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[38] G. Pavlakos, X. Zhou, K. D. G. Konstantinos, and D. Kostas, "Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[39] D. Tomè, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[40] B. Tekin, P. Marquez-neila, M. Salzmann, and P. Fua, "Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation," in *International Conference on Computer Vision*, 2017.

[41] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in *Conference on Computer Vision and Pattern Recognition*, 2014.

[42] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[44] D. King, "Dlib-ml: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[45] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[46] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," *IEEE Transactions on Information Forensics and Security*, 2015.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.