3D Human Body Reconstruction from a Single Image via Volumetric Regression

Aaron S. Jackson[1(✉)], Chris Manafas[2], and Georgios Tzimiropoulos[1]

[1] School of Computer Science, The University of Nottingham, Nottingham, UK
{aaron.jackson,yorgos.tzimiropoulos}@nottingham.ac.uk
[2] 2B3D, Athens, Greece
chris.manafas@2b3dglobal.com

Abstract. This paper proposes the use of an end-to-end Convolutional Neural Network for direct reconstruction of the 3D geometry of humans via volumetric regression. The proposed method does not require the fitting of a shape model and can be trained to work from a variety of input types, whether it be landmarks, images or segmentation masks. Additionally, non-visible parts, either self-occluded or otherwise, are still reconstructed, which is not the case with depth map regression. We present results that show that our method can handle both pose variation and detailed reconstruction given appropriate datasets for training.

Keywords: 3D reconstruction · Human body reconstruction
Volumetric regression · VRN · Single image reconstruction

1 Introduction

3D reconstruction is the process of estimating the 3D geometry from one or more 2D images. In this work, we focus on reconstruction of the human body from a single image, including the non-visible parts which have been self-occluded. Our method builds upon that of [1] where a 3D face is directly regressed from a single image, using what they refer to as a "Volumetric Regression Network" (VRN). In this paper, we show that the same idea can be applied to other deformable objects, in particular, the human body. This poses an array of challenges which are not present when reconstructing the face. While we are still only reconstructing an object of a single class, the body has many more axes of rotation compared to a face. As such, human body reconstruction is often considered to be a very difficult problem (Fig. 1).

Motivation. The pipelines required for 3D human reconstruction (and 3D reconstruction in general) are typically based on solving difficult non-convex optimisation problems. Perhaps the most common approach to 3D human body reconstruction is to fit a shape model. For example, the recent method of [2], uses optimisation to fit a 3D shape model to 2D body joints. However, optimisation methods are sensitive to initialization and are easily trapped to local minima, both of which are exacerbated by occlusions and potential scale changes.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11132, pp. 64–77, 2019.
https://doi.org/10.1007/978-3-030-11018-5_6

Fig. 1. Some example results using our method when trained with a high quality detailed training set.

In this work, we aim to significantly reduce the complexity of standard 3D human reconstruction techniques - to the point where it could just as easily be treated as a segmentation task. We do this by directly regressing a volumetric representation of the 3D geometry using a standard, spatial, CNN architecture, where the regressed volumetric structure is spatially aligned with the input. Notably, we do **not** regress a depth map; the 3D structure is regressed as slices and recovered from its volumetric representation using a standard surface extraction algorithm, such as Marching Cubes [3]. In summary, our main contributions in this work are as follows:

1. We are the first to apply volumetric regression networks [1] to the problem to human body reconstruction, not just human faces.
2. We propose several improvements to the network architecture described in [1], which show significant performance improvements. These include introducing intermediate supervision, using more advanced residual modules and altering the network structure by increasing the number of hourglass modules by reducing the number of residual modules.
3. We show that VRN is capable of reconstructing complex poses when trained on a suitable dataset.
4. Finally, we show that given high quality training data, our method can learn to produce previously unseen, highly detailed, full 3D reconstructions from only a single image. To the best of our knowledge, there is no other method capable of obtaining results with such high fidelity and reliability as ours.

The remainder of the paper is structured as follows: First, a review of closely related work on 3D human body reconstruction and human pose estimation is given. We then describe our method, including the volumetric representation we have already mentioned briefly, followed by the datasets and the training procedure. Next we will discuss several architectural variants of VRN, followed by results from a network trained with pose-variant data, but little detail. Finally, we will show results which have been generated by training a model with highly detailed data.

2 Closely Related Work

In this section we will give an overview of recent and popular approaches to human pose estimation (often a prerequisite to human reconstruction) and 3D reconstruction methods, both working from images and from landmarks.

Human Pose Estimation. All modern approaches to estimating the human pose are based on methods employing CNNs. These methods generally fall into one of two categories. The first is to directly regress the coordinates of the joints using an L2 (or similar) loss [4–10]. In particular, [5] estimate the 3D pose by combining the 2D predictions with image features. An autoencoder is employed in [6] to constrain the pose to something plausible. Similarly, [8] have the same goal but achieve this by using a kinematic model. Synthetic data is used for the full training procedure in [9], to ensure that the network is trained with accurate data. However, in [10], they only augment their existing training set with synthetic data. The second approach to CNN based human pose estimation is to regress a heatmap [11–14]. In [11] they do this from video. In [12] they regress a 3D heatmap, which is a similar idea to our own work. Another temporal based approach is described in [13], where the 2D landmarks are first refined also as a heatmap. A part based heatmap regression approach is shown in [14].

In this work, we do not aim to estimate the human pose as a set of coordinates. Instead, we aim to reconstruct the full 3D geometry of the human, from just a single image. This includes any parts of the body which are self occluded. However, in doing so, we optionally make use of information from a human pose estimation step, which is provided to the network as 16 channels, each with a Guassian centred above the respective landmark.

Reconstruction from Image. Many human reconstruction methods estimate the geometry from one or more images. For example, [15–17] fit a model based on a single RGB or grey scale images. In particular, [16] fit a skeleton model to the image by estimating the scale and pose of each body part separately. In [17], they fit a shape model initialised by a user clicking on separate body parts, assisted by a segmentation mask. Another shape model based approach is proposed in [15], using the SCRAPE model [18], which is fitted with a stochastic optimisation step. A general shape fitting method for reconstruction is proposed in [19], where two Gaussian models are used - one for shape and one for pose, by solving non-linear optimisation problems. The authors demonstrate this method on human bodies and sharks. In [20], a single image and corresponding landmarks are used to lookup a similar human pose using a kd-tree, containing about 4 million examples. A method intended for multi-instance model fitting from a single image is described in [21].

Several methods aim to estimate the 3D geometry using only the landmarks extracted via human pose estimation [2,22]. Particularly, SMPLify [2] (which uses the SMPL model [23]), was extended to also include further guidance from an segmentation mask in [24]. However, such an approach will never be able to capable of regressing finer details, unless information from the image is also captured.

Aside from SCRAPE [18] and SMPL [23], mentioned earlier, Dyna, the shape model capable of capturing large variations in body shape is presented in [25], but without an accompanying fitting method from a single image. A very recent shape model called Total Capture [26] captures many aspects of the body which are typically ignored by other shape models, including the face and hands.

Our work is different from all of the aforementioned methods in that we do not regress parameters for a shape model, nor do we regress the vertices directly. Further more, our method skips the model generation step entirely, which avoids the need to find dense correspondence between all training examples. Instead, we constrain the problem to the spatial domain, and directly regress the 3D structure using spatial convolutions in a CNN, via a volumetric representation from which the full 3D geometry can be recovered.

3 Method

This section describes our proposed method, including the voxelisation and alignment procedures.

3.1 Volumetric Regression

In this work, our goal is to reconstruct the full geometry of a human body from just a single image. There are several ways of estimating the geometry using deep learning. The first is to directly regress the vertices using a top-down network such as VGG [27] or ResNet [28] trained with an L2 loss. This has at least two disadvantages: firstly it requires the training data to be resampled to have a fixed number of vertices, which implies finding correspondence between all vertices of all meshes. Secondly, and more importantly, training a network to directly regress a very large number of vertices is hard. A common, and more efficient alternative is to regress the parameters of a 3D shape model. However, as these parameters are not scaled equally, it is necessary to employ normalisation methods, such as weighting the outputs using the Mahalanobis distance which has been also proven challenging to get it working well [1]. Additionally, 3D shape model based approach are known to be good at capturing the coarse shape but less able at capturing fine details (in the case of detailed 3D reconstruction).

To eliminate the aforementioned learning challenges, we reformulate the problem of 3D reconstruction by constraining it to the spatial domain, using a standard convolutional neural network. Our approach can be thought of as a type of image segmentation where the output is a set of slices capturing the 3D geometry. Hence architecturally one can use standard architectures for (say, semantic) segmentation. Following the work of [1] on human faces, we do this by encoding the geometry of the body in a volumetric representation. In this representation, the 3D space has been discretised with a fixed dimensionality. Space which is *inside* the object is encoded as a voxel with value equal to one. All other space (i.e. background or unknown object classes) are encoded with

a voxel with a value equal to zero. For this particular application, the dimensionality of our volumes are $128 \times 128 \times 128$, which given the level of detail in our training set, is more than adequate (although we show in Sect. 6 results with much greater detail, and only a slightly larger volume). One of the main advantages of this representation is that it allows the non-visible (self-occluded or otherwise) parts of the geometry to also be reconstructed. This is not the case in methods attempting to reconstruct the body using depth map regression.

One of the most important aspects to note in the case of training a volumetric regression network is that the input and output must be spatially aligned. Put simply, the 2D projection of the target object should do a reasonable, if not very good, job at segmenting the input. Through experimentation, we have found that it is possible to ignore spatial alignment, as long as the pose is fixed (i.e. always frontal). However, ignoring spatial alignment will severely impact the performance of the method.

When trained to receive guidance from human pose estimation, landmarks are passed to the network as separate channels, each containing a Gaussian centred over the landmark's location. The Guassians have a diameter of approximately 6 pixels.

3.2 Dataset and Voxelisation

While Human3.6M [29,30] does include its own 3D scans, they are not in correspondence with the video frames. As such, we produced our training data by running SMPLify [2] on the Human3.6M dataset. The landmarks required by SMPLify were generated using the code made available with [14]. The fitted meshes were voxelised at a resolution of $128 \times 128 \times 128$. In terms of depth, the meshes are first aligned by the mean of the Z component. However, through experimentation, we found that as long as the Z alignment is performed in a seemingly sensible way, and remains stable across all images, the network will learn to regress the 3D structure without issue. Random scale augmentation was performed in advance of the training procedure, as doing this on-the-fly (for 3D volumes) can be quite demanding in terms of CPU usage.

An unfortunate side effect of using SMPLify to generate our training data is that it is not possible to regress features such as fingers or facial expressions. SMPLify does not model these, and as such, their pose remains fixed across all images. It also becomes a bottleneck in terms of performance. We show in Sect. 6, using a different dataset, that very high quality reconstruction is also possible with our proposed method.

3.3 Training

Our end-to-end network was trained using RMSProp [31] optimisation with a learning rate of 10^{-4}, which was reduced after approximately 20 epochs to 10^{-5} for 40 epochs. We did not observe any performance improvement by reducing this learning rate further. A batch size of 6 was used across 2 NVIDIA 1080 Ti graphics cards. During the voxelisation, random scale augmentation was applied.

Applying scale augmentation to a 3D volume on the fly, is very CPU intensive and slows down the training procedure too much. During training, augmentation to the input image was applied. This on-the-fly augmentation included colour channel scaling, random translation and random horizontal flipping.

4 Architecture

In the following subsections, we introduce the several architectural options we have explored as extensions to [1]. Our first network is the same as the one used in [1], referred to as *VRN - Guided*, which establishes our baseline. This network employs two Encoder-Decoder ("hourglass") networks in a stack. We follow a similar design, aside for the changes described in this section. All of our architectures were trained with the same loss function as in [1]:

$$l_1 = \sum_{w=1}^{W} \sum_{h=1}^{H} \sum_{d=1}^{D} [V_{whd} \log \widehat{V}_{whd} + (1 - V_{whd}) \log(1 - \widehat{V}_{whd})], \tag{1}$$

where $\widehat{V}_{whd}$ is the corresponding sigmoid output at voxel $\{w, h, d\}$.

4.1 Ours - Multistack

This network makes the following changes to the *VRN - Guided* baseline network. We half the number of residual modules from four to two. In doing so, we also halved the memory requirements, allowing us to increase the number of hourglass modules in the stack, from two to four. Next, we replace the original residual module used in *VRN - Guided* with the multi-scale residual module proposed in [32]. We also show the performance improvement from introducing just this component in the results section. Finally, we introduce supervision after each hourglass module. We therefore have four losses. Each hourglass module forks to provide features for the next hourglass, and to regress the volumetric representation. The performance after each hourglass improves. We found that there was no benefit to adding more than four hourglass networks as the performance just fluctuates as more are added. This network is depicted in Fig. 2.

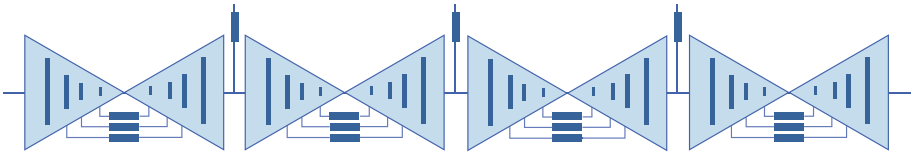

Fig. 2. The *Ours - Multistack* network. Dark blue boxes represent residual modules. Each Encoder-Decoder module has its own loss, while still passing features to the next module. (Color figure online)

4.2 Ours - Image Only

Our standard network (*Ours - Multistack*) is trained to receive guidance from the landmarks, while also using useful information from the images. With this network, we try to measure the impact of training with just images, while keeping the architecture identical. We call this network *Ours - Image Only*. We expect that the performance of this network be significantly lower than when guidance from the human pose is also provided.

4.3 Ours - Landmarks Only

Many methods, such as [2,22], use only the landmarks as input during training and inference. Hence, it is an interesting investigation to measure the performance of our method when only landmarks are provided, without the image. As such, we trained *Ours - Landmarks Only*. However, using only landmarks to fit a shape model results in generic appearing fittings. Provided high quality training data is available, our method can regress these fine details and match the body shape when also provided with the image.

4.4 Ours - Mask Only

Our method does not rely on a segmentation mask, as is the case in [33]. However, there is no reason why our method cannot reconstruct 3D geometry from a single segmentation mask, or silhouette. To show this, we train another network, *Ours - Segmentation Mask* which accepts only a single channel, containing the mask of the target object. From this, the network reconstructs the 3D geometry in the same way. Once again, this network has an identical configuration to *Ours - Multistack*, apart from the first layer receiving a different number of inputs. We expect this network to perform quite well since the segmentation mask we are providing to the network is the projection of the target volume.

4.5 Ours - 3D Convolution

While volumetric CNNs can likely outdo a spatial network in terms of performance, on this task, the memory requirements are much higher than that of a spatial CNN. So much so, that employing volumetric CNNs at a suitable resolution is not currently possible. However, we were interested to test a compromise between the two and train a volumetric CNN where the filters are flat. More concretely, where f is the number of features, our filters had sizes $f \times 3 \times 1 \times 1$, $f \times 1 \times 3 \times 1$ or $f \times 1 \times 1 \times 3$. These were combined into a flat volumetric residual module, as shown in Fig. 3, heavily inspired by [34]. This network also takes as input the image with corresponding landmarks. To provide a fair comparison with the other methods, we match the number of floating point operations of this network to *Ours - Multistack* by reducing the number of parameters (which also allows the network to fit into memory).

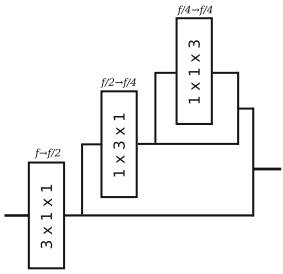

Fig. 3. A "flat" volumetric residual block

5 Results

In this section we will give an overview of the performance of the architectures we have described above. For each network, we give our results as an Intersection over Union (IoU) score, which is defined as the number of intersecting set voxels over the number voxels set in either volume. These numeric results may be found in Table 1. We will discuss these results in more detail in the proceeding paragraphs.

We show visual results for *Ours - Multistack* in Fig. 4. The quantitative results suggest that the changes we made to the baseline network *VRN - Guided* helped quite significantly, offering a performance increase of over 4% in terms of IoU. From this performance improvement, more than 2% was due to using the residual module proposed in [32], this can be seen from the results for *Ours - Old Residual*. As our data is generated by SMPlify [2], we are unable to provide a quantitative comparison with this method.

As expected, removing either the landmarks *or* the image reduces performance. The best performance is attainable by providing the network with both the image and landmarks, as seen quantitatively between *Ours - Multistack*, *Ours - Landmarks Only* and *Ours - Image Only*. Also unsurprisingly, landmarks alone offers better performance than the image alone. This is true at least in this case, as the groundtruth model has no detail. We also show performance where only the segmentation mask is provided to the network (this is not provided in the case of *Ours - Multistack*). These results are labelled *Ours - Mask Only*. We expected this network to perform better than the landmarks or image only networks, as the mask we provided was a direct 2D projection of the target volume.

Notes on Performance. A single forward pass through our network takes approximately 200 ms on an NVIDIA 1080 Ti GPU. This produces the volumetric representation. Surface extraction introduces 200–600 ms overhead depending on the implementation used. Significantly higher performance may be achieved with smaller volumes, but this will result in a lower level of detail. Training typically takes about two days.

Fig. 4. Visual results from our main network, *Ours - Multistack*, on a test split of Human3.6m [30]. These results demonstrate VRN's ability to deal with large and complex pose. We also show the reconstructions with the texture projected onto them.

Table 1. Numerical performances of our proposed method and additional architectural experiments, all on data generated using SMPLify.

Method	IoU @ epoch 30	IoU @ epoch 60
VRN - Guided (Baseline)	61.6%	63.9%
Ours - Multistack	61.1%	**68.3%**
Ours - Old Residual	60.5%	66.1%
Ours - Landmarks Only	58.6%	61.0%
Ours - Image Only	46.8%	48.3%
Ours - Mask Only	52.8%	53.0%
Ours - 3D Convolution	57.3%	61.6%

6 High Quality Training Data

In the previous section, we showed that our method can reconstruct bodies of very large pose. However, due to the dataset we trained on, we are only able to regress the coarse geometry without any detail. Detailed 3D reconstruction was also not demonstrated in the case of faces in [1], which was also due to the lack of a detailed dataset. Hence, in this section, we demonstrate that VRN *is* capable of regressing details when a high quality dataset is provided. For this experiment, we use our best performing network *Ours - Multistack*.

Our dataset consists of highly detailed 3D scans from 40 participants, 4 of which were reserved for quantitative testing, but all of which are quite restricted in terms of pose. Only one scan per participant was available. These models do not have a corresponding image which is aligned with the model. As such, we rendered and voxelised these models under a large variety of different lighting conditions, scales and views to create our training set consisting of approximately 20,000 samples which are spatially aligned. The voxelisation was performed at a resolution of $128 \times 256 \times 96$, which efficiently encapsulates the poses found in the dataset. As in our previous experiment, the Z alignment was performed by the mean Z component. Unfortunately we are not able to publicly release this dataset.

6.1 Performance

The four models which we reserved for testing were also rendered and voxelised in the same way as above, to produce 60 testing images. Our method reconstructs these with an IoU of 78%. This is significantly higher than the reconstructions in our previous experiment. This is likely due to the better spatial alignment between the training images and target. Additionally, we show qualitative results on real-world images taken from the web[1]. These reconstructions can be found in Fig. 5. We show the backsides of these reconstructions, which demonstrate the networks ability to reconstruct the self-occluded body parts. Finer details can be seen in the wrinkles of clothing. As our method was trained on synthetic data, we believe there may be some performance degradation on real-world images. Additionally, several of the poses found in the reconstructions in Fig. 5 are not found in the 36 training samples. This suggests that VRN is somewhat tolerant to previously unseen poses.

[1] These images are licensed under Creative Commons. Attribution, where required, will be provided on our website.

Fig. 5. Example 3D reconstructions from the web (Creative Commons) using our method trained with high quality training data. The first row shows the input image, the second shows the 3D reconstruction from the front, and the third row shows the 3D reconstruction when views from behind (i.e. the hallucinated side, in the case of these images). The final row shows the frontal reconstruction with the projected texture. These results show that VRN is capable of regressing finer details.

7 Conclusion

In this work we have shown that using Volumetric Regression Networks, as described in [1], for the task of 3D reconstruction, is not restricted to the simpler task of face reconstruction. Nor is it a limiting factor in terms of detail, despite the small size of the volumes we are working with. We have proposed several improvements to the original VRN which improve the performance quite substantially. Finally, we have shown, by using two different datasets, that VRN can regress both unusual poses (in networks trained on Human3.6m), and high levels of detail (in the case of our private but detailed dataset). We believe that given a large enough dataset containing many pose variations, and high levels of detail, the network will be capable of large pose 3D human reconstruction, while also capturing fine details, from a single image.

Acknowledgements. Aaron Jackson is funded by a PhD scholarship from the University of Nottingham. Thank you to Chris Manafas and his team at 2B3D for providing data for the experiments. We are grateful for access to the University of Nottingham High Performance Computing Facility, which was used for data voxelisation.

References

1. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1031–1039. IEEE (2017)
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep It SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part V. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
3. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: ACM SIGGRAPH Computer Graphics, vol. 21, pp. 163–169. ACM (1987)
4. Li, S., Chan, A.B.: 3D human pose estimation from monocular images with deep convolutional neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9004, pp. 332–347. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16808-1_23
5. Park, S., Hwang, J., Kwak, N.: 3D human pose estimation using convolutional neural networks with 2D pose information. In: Hua, G., Jégou, H. (eds.) ECCV 2016, Part III. LNCS, vol. 9915, pp. 156–169. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_15
6. Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., Fua, P.: Structured prediction of 3D human pose with deep neural networks. arXiv preprint arXiv:1605.05180 (2016)
7. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3D body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 991–1000 (2016)

8. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., Jégou, H. (eds.) ECCV 2016, Part III. LNCS, vol. 9915, pp. 186–201. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_17
9. Chen, W., et al.: Synthesizing training images for boosting human 3D pose estimation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 479–488. IEEE (2016)
10. Ghezelghieh, M.F., Kasturi, R., Sarkar, S.: Learning camera viewpoint using CNN to improve 3D body pose estimation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 685–693. IEEE (2016)
11. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4966–4975 (2016)
12. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1263–1272. IEEE (2017)
13. Mehta, D.: Vnect: Real-time 3D human pose estimation with a single RGB camera. ACM Trans. Graph. (TOG) **36**(4), 44 (2017)
14. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VII. LNCS, vol. 9911, pp. 717–732. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_44
15. Balan, A.O., Sigal, L., Black, M.J., Davis, J.E., Haussecker, H.W.: Detailed human shape and pose from images. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
16. Grest, D., Herzog, D., Koch, R.: Human model fitting from monocular posture images
17. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1381–1388. IEEE (2009)
18. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM Transactions on Graphics (TOG), vol. 24, pp. 408–416. ACM (2005)
19. Chen, Y., Kim, T.-K., Cipolla, R.: Inferring 3D shapes and deformations from single views. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 300–313. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15558-1_22
20. Jiang, H.: 3D human pose reconstruction using millions of exemplars. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 1674–1677. IEEE (2010)
21. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
22. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3D human pose from 2D image landmarks. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 573–586. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_41
23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248 (2015)

24. Varol, G., et al.: Learning from synthetic humans. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), pp. 4627–4635. IEEE (2017)
25. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: a model of dynamic human shape in motion. ACM Trans. Graph. **34**(4), 120:1–120:14 (2015)
26. Joo, H., Simon, T., Sheikh, Y.: Total capture: a 3D deformation model for tracking faces, hands, and bodies. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
29. Ionescu, C., Li, F., Sminchisescu, C.: Latent structured models for human pose estimation. In: International Conference on Computer Vision (2011)
30. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
31. Hinton, G., Srivastava, N., Swersky, K.: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent
32. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: International Conference on Computer Vision (2017)
33. Sigal, L., Balan, A., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: Advances in Neural Information Processing Systems, pp. 1337–1344 (2008)
34. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5534–5542. IEEE (2017)