Deep 3D Human Pose Estimation

Ossama Ahmed
ETH Zurich
oahmed@student.ethz.ch

Guillem-Torrente Martí
ETH Zurich
tguillem@student.ethz.ch

ABSTRACT

Recent advances in scientific computing hardware and the increased availability of public datasets have allowed deep learning to skyrocket the performance of the state-of-the-art models for the problem of 3D Human Pose Estimation. We propose a pipeline that comprises one of the most recent approaches, namely the High-Resolution Network, combined with a low-weight baseline model for extracting the 3D skeleton of human subjects of the Human3.6M dataset. Our approach splits the challenges of the task in image-related, and geometric-related, where each group is processed by a model specialized in one of the two. We show that our model achieves good results despite a relatively low training time (around 24 hours), although it still has some trouble at discriminating some of the upper and lower limbs. We finally propose further training strategies to help the model deal with its current limitations.

1 INTRODUCTION

The 3D Human Pose Estimation problem consists on regressing the 3D position of a set of different anatomical structures of the human body (typically some joints) out of a sequence of images or videos. This challenge has been tackled numerous times in the computer vision research field, as it is of great interest due to the diversity of applications, including AR, VR, CCTV surveillance, character animation or cinematography among others [9, 13].

In the past, some of the approaches to tackle this problem included human fine-tuned algorithms such as pictorial representations, where an optimization algorithm would find the best fit of a graphical model to the image. This approach was in fact first introduced back in 1973 by [4] and later on refined by studies such as [1] with good success back at the time. Currently however, improvements in hardware and the increasing availability of human pose datasets have pushed even further the state-of-the-art performance by leveraging deep learning and especially convolutional neural networks [14, 16, 17].

In this study, we re-implement one of the most recent state-of-the-art deep learning pipelines, namely the High Resolution Network (HRNet) [12] proposed in 2019, and re-train it on a subset of the Human3.6M dataset [5], where subjects are discretized using 17 joints.

2 APPROACH

In this section we cover our approach to tackle the 3D human pose challenge using deep neural models.

Executive Summary

Our pipeline is composed of two neural networks concatenated in series:

- A deep convolutional model receives images as inputs and outputs 17 2D heatmaps of size $[64 \times 64]$ (one per joint) per image. We use HRNet [12] for this part.
- A fully connected model with dropout units and skip connections maps the 2D joint coordinates to the 3D space. We use the model proposed in [7].

We augment the dataset with image-domain pixel transformations (brightness, hue, contrast, saturation and Gaussian noising), and with data rotation and translation. The mapping between 2D heatmap to 2D coordinates is performed by means of *argmax* sampling, and the optimal parameters for both models are obtained though L2 loss gradient descent.

2.1 Overview

We propose a pipeline that splits the challenges of the 3D human pose estimation task into two semantically different domains, namely the 2D image space and the 3D space. We exploit several features of neural network design to obtain two specialized regressors in one of the two contexts. We use a two-model pipeline where the first one infers the 2D position of the joints in the image space from the input image, and the second one projects these 2D predictions to 3D coordinates. In fact, a similar approach has been used in past researches such as [2, 7, 18]. In particular, in [7] the authors empirically show that the 2D to 3D mapping can be performed with competitive results without any image information whatsoever, which suggests that the 3D skeleton representation has in fact a reliable latent space in 2D.

Image to 2D coordinates.: With this setup, the first regression model must learn how to deal with image-related challenges such as occluded parts of the body and non-visible joints, left-right side differentiation, clothing and lighting changes, position in the image and posture. Notice that all the information to overcome these is self-contained in the image itself. Because of the necessity of invariance or equivariance throughout all these sources of variability, as indicated by [6], we exploit convolutional layers, and data augmentation. In particular, we use:

- Pixel color augmentations to handle the differences in clothing (hue, contrast, saturation).
- Brightness augmentation for illumination.
- Gaussian noising to avoid over-fitting to high resolution features (such as for instance the infrared fiducials worn by the subjects used to generate the ground truth data)
- Translation and rotation (in multiples of 90 degrees) augmentation for their respective invariance.

In our particular case, the input image $\mathbf{I}$ of fixed dimensions $[255, 255, 3]$ is fed through a deep convolutional model that outputs 17 heatmaps of sizes $[64 \times 64]$. Each heatmap $\mathbf{H}_i,\ i \in \{0, \ldots, 16\}$ is in fact a down-scaled representation of the original input, where each pixel in $\mathbf{H}_i$ represents the probability of joint i being in the corresponding pixel of I. We use the HRNet introduced by [12] for this first task.

Then, the heatmaps are converted back to joint 2D coordinates in the image space by *argmax* sampling. This operation is not differentiable, so the model must be trained using heatmaps as ground truth instead of the actual 2D coordinates.

2D to 3D coordinates.: Last but not least, the second model is in charge of learning to infer 3D structure from the 2D image domain, which is a geometrical problem, semantically different from the previous challenges.

The 2D pixel coordinates are projected to 3D coordinates (where the hip joint is the origin of the frame), through a simple linear regression model introduced in [7]. This model essentially consists on repetitions in series of a main building block consisting on a linear layer, a batch normalization layer, a ReLU activation, a dropout unit and a skip connection. We describe with more detail the two specified components in the next two sections.

Evaluation metrics.: Finally, we use the Mean Per-Joint Position Error (MPJPE) metric to numerically assess the performance of our pipeline. Given the number of joints N, and the ground-truth and predicted positions $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^{3N}$ this value is computed by (1).

$$MPJPE(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N-1} \sum_{n=0}^{N} \|\mathbf{x_n} - \hat{\mathbf{x}}_n\|_2^2 \tag{1}$$

2.2 High-Resolution Net

Typical state-of-the-art convolutional models for pose estimation (both in 2D or 3D) are based on concatenated high-to-low and low-to-high resolution building blocks [12]. The motivation of this approach is to use the high-to-low module to extract a high-level representation of the image and the low-to-high to gain back some of the lost resolution for improved prediction precision [3, 10]. Some examples of this include the popular Hourglass Network [10, 11], which uses this same approach with added skip connections to help both keep high resolution information and help propagate the gradients in training time.

The motivation behind the HRNet model is to extract global, context-aware information, while also keeping the resolution of the intermediate channels high, so as not to loose spatial precision. However, instead of using multi-resolution layers stacked in series like in [10, 11], the HRNet works by keeping models operating at different resolutions in parallel, repeatedly fusing the information flowing through each one by means of an *information exchange* unit. The advantage of having multiple resolution data channels in parallel is that the network is never forced to re-construct a high resolution image from a low resolution one, potentially keeping better the spatial information of the original image.

Furthermore, recent studies have shown that reformulating the supervision of this model from 2D coordinates to 2D heatmaps over the entire output space acts as a *natural tie* between the image and the position of the different joints [8, 15]. This not only helps in training time by providing a richer loss function, but it also has shown to help the model learn the graphical structure of the human body [15]. Indeed, the HRNet also adopts this heatmap supervision technique, where the ground truth 2D annotations are manually transformed to this heatmap representation. Given the function $\mathbf{G}(\cdot)$ that maps a pose $\mathbf{x} \in \mathbb{R}^2$ to a heatmap of size 64×64, and a prediction $\hat{\mathbf{H}}$ of the same shape provided by the deep model, the loss function of the can be written as in (2).

$$\mathcal{L} = \frac{1}{NZ} \sum_{n=0}^{N-1} \sum \left\|\mathbf{G}\left(\mathbf{x}_n\right) - \hat{\mathbf{H}}_n\right\|_2^2 \tag{2}$$

Given $N = 17$ joints and the heatmap size $Z = 64 \cdot 64 = 4096$. Finally, also following the procedure suggested in [12], the heatmaps are converted back to 2D coordinates by means of *argmax* sampling as a post-processing step.

2.3 2D to 3D mapping

The second part of the pipeline is has the task of *lifting* the joint coordinates from the 2D image space to the 3D space. In order to simplify the problem, a typical approach proposed by multiple datasets is to use the origin of coordinates to be the hip joint. This strategy effectively reduces the number of joint coordinates to be predicted from 17 to 16, and is also adopted by the Human3.6M dataset.

For this task, we leverage the approach proposed in [7], in which a *simple yet effective* dense feed-forward network is used. In the original report, the authors claim that this model architecture is capable of getting state-of-the-art results in 3D pose estimation using an *off-the-shelf* pre-trained 2D regressor as input. Furthermore, it has the advantage that due to its simplicity in structure and low number of parameters (4.3M), it can be trained very fast.

The main building block of the network are two repetitions of the sequence: Linear layer with 1024 units → Batch normalization → ReLU activation → Dropout with 0.5 rate. A skip connection is added from the beginning until the end of the block, and as suggested by the original article, the entire model is composed of two blocks. We empirically confirm that adding more blocks or modifying the dropout rate does not yield any significant improvement on the performance of the model.

furthermore, in [7], both the 2D inputs and the 3D outputs are normalized during a pre-processing step. This normalization includes subtracting to every joint in the input and output the respective average joint position over the entire training dataset, and dividing that by the standard deviation. Thanks to this pre-processing, and to the regularization added by the dropout and the batch normalization layers, plus the translation and rotation data augmentation, this model is capable of generalizing the lifting task without much risk of over-fitting to the training data.

2.4 Training specifications

Both models are trained on a subset of the Human3.6M dataset consisting on a total of 312188 labeled images with 2D and 3D joint coordinates. Due to its structure and complexity, the HRNet model is trained on an Nvidia GTX 1080 GPU. Using a batch size of 32. One epoch over the dataset takes approximately 2:15 hours to complete, and this model is trained for a total of 12 epochs (27-30 hours). On the other hand, the 2D-to-3D model is much more lightweight, so we train it on an Xeon E5-2697v4 CPU, where one epoch with a batch size of 512 takes around 8 minutes to complete. This model is trained for a total of 180 epochs (24h).

In both cases, the training is implemented with the Tensorflow 1.12 framework, and the optimizers used are Momentum Optimizers using a momentum parameter of 0.9 and a learning rate of 0.003.

3 RESULTS

All in all, the proposed pipeline achieves 105 MPJPE score on the entire test set provided. We show some results for images of the training and test sets in Figures 1 and 2 respectively. We can see that the HRNet model predicts with quite good accuracy the heatmaps on the corresponding joints on the training set despite all the augmentations, which leads to good predictions of the 2D joint coordinates in columns 3 and 4 from Figure 1. In fact, the final loss after the 27 hours of training time is of 0.17. The lifting of 2D data onto the 3D space is also capturing most of the depth information, although it is clearly the main source of error in the training images. We also observed that some of the joints were easier for the HRNet to detect early on during training; such as the neck or the feet. On the other hand, we noticed that some joints such as the knees, elbows or the hands are harder to properly detect since they require the network to learn more about the human structure. In particular, these joints also have the added difficulty that they come in pairs (left-right) and the network needs to know which half of the body is which.

The testing data from Figure 2 in fact validates this hypothesis. In the lower row of images we can see that the two knees are merged as the network has confused them, and furthermore that it still does not completely know how to deal with occlusions, as the hidden right wrist is placed on top of the left. Notice, however, that the 2D-to-3D regressor is able to properly infer that the right wrist should be at a different depth than the left, although the bending angles of the knees look a bit exaggerated.

4 CONCLUSION

In this study we presented a two-stage pipeline to tackle the human 3D pose estimation problem, composed by two state-of-the-art deep models. We showed that the HRNet combined with a simple baseline regressor produces concise 3D pose estimate even with relatively small training times (around 24 hours). Our pipeline leverages the recent deep learning methods to implicitly learn a spatial model of the human body skeleton without the need for graphical-model inference. As a result, the pipeline achieved better results than the off-the-shelf RESNET-50 on a subset of the Human3.6M dataset. For future work, having supervision at different levels of the HRNet might be an interesting venue to explore. Furthermore, it would be very interesting to apply some augmentation strategy to help the model learn how to deal with occlusions, like for instance artificially covering specific parts of the body during training time, particularly the joints that have left and right pairs.

REFERENCES

[1] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 2014. 3D Pictorial Structures for Multiple Human Pose Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 38. https://doi.org/10.1109/CVPR.2014.216

[2] Ching-Hang Chen and Deva Ramanan. 2017. 3D Human Pose Estimation = 2D Pose Estimation + Matching. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5759–5767.

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2017. Cascaded Pyramid Network for Multi-Person Pose Estimation. *CoRR* abs/1711.07319 (2017). arXiv:1711.07319 http://arxiv.org/abs/1711.07319

[4] M. A. Fischler and R. A. Elschlager. 1973. The Representation and Matching of Pictorial Structures. *IEEE Trans. Comput.* 22, 1 (Jan. 1973), 67–92. https://doi.org/10.1109/T-C.1973.223602

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

[6] Eric Kauderer-Abrams. 2018. Quantifying Translation-Invariance in Convolutional Neural Networks. *CoRR* abs/1801.01450 (2018). arXiv:1801.01450 http://arxiv.org/abs/1801.01450

[7] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A simple yet effective baseline for 3d human pose estimation. *CoRR* abs/1705.03098 (2017). arXiv:1705.03098 http://arxiv.org/abs/1705.03098

[8] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *CoRR* abs/1705.01583 (2017). arXiv:1705.01583 http://arxiv.org/abs/1705.01583

[9] Tobias Nägeli, Samuel Oberholzer, Silvan Plüss, Javier Alonso-Mora, and Otmar Hilliges. 2018. Real-time Environment-independent Multi-view Human Pose Estimation with Aerial Vehicles. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH ASIA)*.

[10] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. *CoRR* abs/1603.06937 (2016). arXiv:1603.06937 http://arxiv.org/abs/1603.06937

[11] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *CVPR*.

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *CoRR* abs/1902.09212 (2019). arXiv:1902.09212 http://arxiv.org/abs/1902.09212

[13] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Efficient Object Localization Using Convolutional Networks. *CoRR* abs/1411.4280 (2014). arXiv:1411.4280 http://arxiv.org/abs/1411.4280

[14] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2015. Efficient object localization using Convolutional Networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 648–656.

[15] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *CoRR* abs/1406.2984 (2014). arXiv:1406.2984 http://arxiv.org/abs/1406.2984

[16] Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, Washington, DC, USA, 1653–1660. https://doi.org/10.1109/CVPR.2014.214

[17] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 4724–4732.

[18] Xingyi Zhou, Qi-Xing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 398–407.

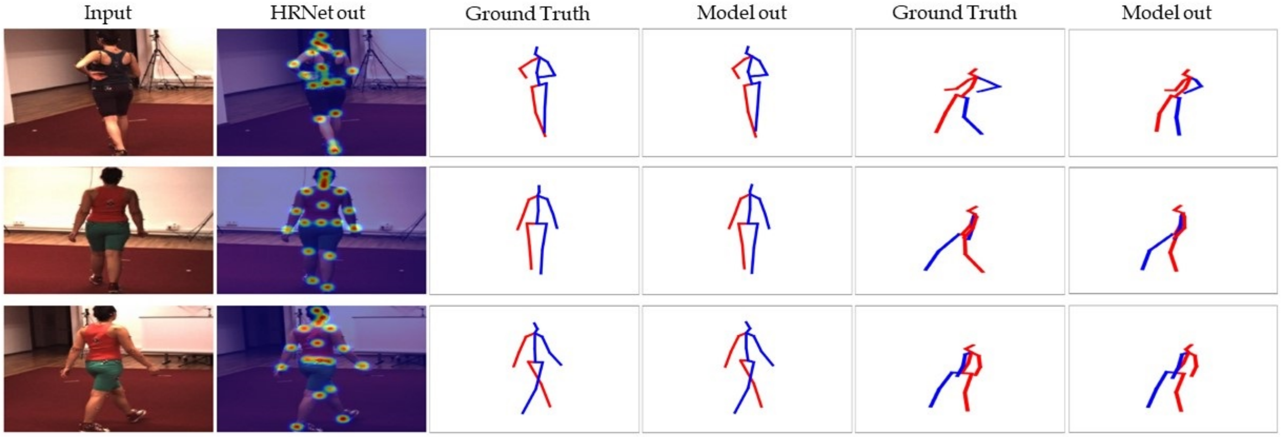

Figure 1: Sample predictions of our pipeline in the training set. First column: the input image. Second column: the overlaid heatmap activations of the HRNet onto the image. Third and Fourth columns: frontal projections of the ground truth and predicted 2D poses. Fifth and Sixth columns: lateral projections of the ground truth and predicted 3D

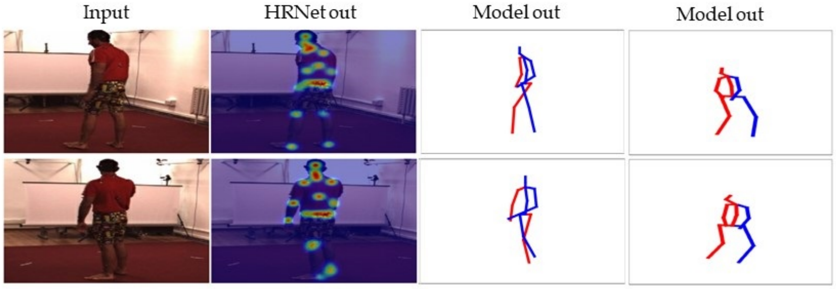

Figure 2: Sample predictions of our pipeline in the testing set (no ground truth available). First column: the input image. Second column: the overlaid heatmap activations of the HRNet onto the image. Fourth column: frontal projection of the predicted 2D pose. Sixth columns: lateral projection of the predicted 3D