3D Human Pose Estimation from RGB Images

ETH Zurich, Machine Perception, Spring Semester 2021

Martin Bucher
bucherma@student.ethz.ch

Valentin Weiss
weissva@student.ethz.ch

Andrin Bertschi
bandrin@student.ethz.ch

ABSTRACT

This work addresses the problem of 3D human pose estimation given solely monocular RGB images and is part of the Machine Perception course at ETH Zurich during Spring Semester 2021. Given a 2D RGB image of a human body, the goal for project 2 was to predict the 3D coordinates of 17 body joints. A preprocessed version of the Human3.6M dataset [3] was provided for training and validation with ground-truth 3D pose annotations, as well as MPII [1], a dataset which contains in-the-wild human images with 2D pose annotations. We explored several different paths and methods making usage of Deep Learning and ran various experiments to validate the different approaches. For an additional comparison, we compare our results to a method exploiting the multi-view property of the Human3.6M dataset which was not allowed as a valid submission, but further shows the potential of such a method. With single monocular images we achieved a final MPJPE of 60.82mm, whereas by using the multi-view property we achieved 37.65 mm. These preliminary results suggest that multi-view geometry can significantly improve the accuracy for human pose estimation.

1 INTRODUCTION

Detecting 3D human poses from static RGB images has received increased attention in many fields including robotics, human computer interaction and autonomous driving and has many applications in a variety of domains. Thanks to the advent of neural networks, especially Convolutional Neural Networks (CNNs), significant progress has been made in recent years in improving the estimation of 3D poses from RGB images.

Existing methods in this area are either regression based or detection based. Detection based methods try to predict a likelihood heatmap for each joint and localize the joint as the point with the maximum likelihood in that map. Such likelihood maps are often referred to as "heatmaps" and there exist methods for both 2D and 3D heatmap estimation. These heatmaps have often a very low resolution (i.e. 64x64 or 80x80), which leads to a certain amount of quantization error for the prediction. Using heatmaps with higher resolution helps to increase the accuracy, but it is computationally heavier and requires more memory, especially for 3D heatmaps.

Human pose estimation in this setting is essentially a regression problem, as we try to regress the 3D locations of certain joints in the euclidean space given only raw image pixels from an RGB image. In order to infer the 3D pose from an image, the method should be invariant to a number of different factors, including background noise, variations in lighting condition and human body shape and size, different clothing textures, skin color, and different degrees of image noise, among many other factors. Further, the human body has a tremendous number of different positions it can take on, which makes 3D pose estimation in-the-wild a very hard task in the domain of Computer Vision.

This work is structured as follows: First, we introduce the different methods we looked at and implemented in order to be able to evaluate their performance. Next, we compare the different methods based on MPJPE, a widely used metric which reports the root-aligned mean per joint position error in millimeter. We conclude with a discussion of the respective performances and final remarks with our findings for this problem setting.

2 METHODS

We explored several methods as part of this project, which led us to different approaches for the final 3D estimation problem. The different methods are briefly discussed in this section.

2.1 Two-Stage Pipeline

In this approach, we first estimate 2D poses as an intermediate step. The 2D poses are learned separately in a supervised manner by a neural network together with its ground-truth 2D labels (in pixel space). The predicted 2D poses are then used to train a second neural network to "lift" the 2D predictions to 3D by estimating the depth of each joint given its 2D prediction. The results are then back projected to the camera or world space.

2.1.1 2D Predictions. We have explored several different architectures to tackle this task. To obtain a 2D prediction, a "Stacked Hourglass" architecture as proposed by Newell et al. [5] showed the best performance overall and was used by many recent architectures focusing on the uplifting part from 2D to 3D space [2, 9, 10]. The Stacked Hourglass (SH) model consists of multiple "hourglass" modules concatenated together, where each hourglass module consists of multiple convolutional and max pooling layers to bring the features down to a low resolution (the bottleneck layer), followed by layers that upsample the low resolution latent representation up to the original input size. Additionally, at the layers with the same resolution, skip connections are used to better preserve features at a given scale and improve the learning phase. The hourglass module as introduced by [5] is shown in Figure 1.

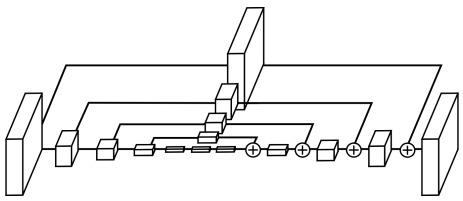

Figure 1: Illustration of an hourglass module as proposed by the Stacked Hourglass (SH) network of [5]

The hourglass network is designed to capture features of the body at different scales and simultaneously develops an understanding of the full body and the relationship between the body parts. It

outputs a 64 x 64 heatmap for each joint containing an estimate for the joint's 2D position within a patch of the input image, marking the likelihood map for a certain joint location. For the heatmaps, a Mean-Squared Error (MSE) loss is applied comparing the predicted heatmap to its ground-truth heatmap consisting of a 2D Gaussian (with standard deviation of 1 px) centered around the joint location. For the input resolution for this network we opted for a 256x256 pixel patch centered around the human subject. This center is given by the metadata of the Human3.6M and MPII datasets [1, 3]. To get an estimation of the 2D coordinates in pixel space, the `argmax` of the likelihood map is taken. We explored the stacking of different numbers of Hourglass modules and found $N = 2$ and $N = 8$ most effective for our setting.

2.1.2 3D Predictions. Given a set of 17 heatmaps containing 2D predictions of each joint, the task here is to "lift" these 2D predictions to 3D by estimating the depth of the joint location in 3D space. This task aims to learn a function $f : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{3n}$ for $n = 17$ joints in our case. We implemented the 2D to 3D uplifting as outlined in the work of Martinez et al. [4], making use of a simple neural network architecture with two subsequent blocks of a dense linear layer with dimension 1024 followed by batch normalization, a ReLU activation, and a dropout layer. This simple architecture performs surprisingly well for this 2D to 3D prediction task. Each joint coordinate is first normalized across the entire dataset by subtracting the mean and dividing by the standard deviation of the data distribution of that specific joint coordinate. We hence normalize all coordinates independently, feed it through the network, and denormalize the joints before performing a subsequent validation or test phase. We experimented with both an L1 and L2 loss over the Mean-Squared Error (MSE) of the 3D predictions.

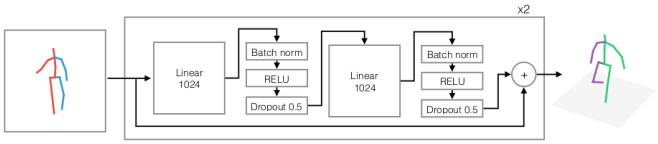

Figure 2: Simple Baseline (SB): 2D to 3D uplifting neural network as proposed by Martinez et al. [4]

2.2 Leveraging Multi-View Geometry

The provided H36m dataset is curated such that we have four image correspondences for every capture of a human subject performing a certain action. This property can be leveraged by applying basic knowledge about multi-view geometry to further guide the training process in the 2D heatmap prediction stage. Our attempt for this is derived from the Cross-View neural network from the work of Qiu et al. [6], where they feed different camera views of the same person into a CNN architecture to obtain four 2D heatmap predictions, belonging semantically together. They further introduce a *Fusion Layer* to fuse different heatmap correspondences together, which increases the performance of the 2D predictions. We implemented their approach for our own heatmap prediction pipeline, hoping to improve the precision of our 2D predictions compared to the Stacked Hourglass prediction, especially as they also experimented by using 80x80 heatmap resolutions compared to 64x64 resolutions. Further, the whole heatmap prediction network is end-to-end learnable.

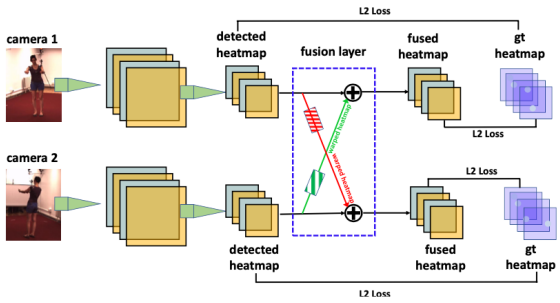

Figure 3: Cross-view fusion for 2D pose estimation as introduced by Qiu et al. [6]

Their method uses a ResNet-152, pretrained on ImageNet, as its backbone and the input image size is either 320x320 or 256x256 pixels. For the 3D uplifting, the authors introduce a Recursive Pictorial Structure Model (RPSM) to recover the 3D pose from multiple 2D poses of the same scene. RPSM is able to dramatically reduce the average error for the joints predictions, and although we would not be allowed to use this method we thought it would be an interesting avenue to approach in order to compare it to other methods for this project. This method is stronger for joints which are occluded from one perspective, but visible from others, which is often the case for the wrist joints.

2.3 Integral Regression

An interesting approach tries to bridge the gap between heatmap representation and joint location regression and was proposed by Sun et al. [8]. Most works (including the previously introduced ones) are either detection based (i.e. heatmaps) or regression based (hence predicting the 3D coordinates directly). The predictions based on heatmaps usually take the `argmax` of a given probability map, where the confidence is the highest. This `argmax` has the drawback that its function is not differentiable. By taking the expectation value of the probability map instead of its maximum (also called `softmax` in other literature), the joint is estimated as the integration of all locations in the heatmap. A joint is hence estimated as the integration of all locations in the heatmap, weighted by their (normalized) probabilities. In discrete space, this can be expressed as:

$$J_k = \sum_{p_z=1}^{D} \sum_{p_y=1}^{H} \sum_{p_x=1}^{W} p * H_k(p)$$

Where D, H, and W are the dimensions of the heatmap H_k (i.e. its resolution) for the joint J_k and $H_k(p)$ denoting the probability of of the joint being at location p. We further denote this method as IR in the subsequent text.

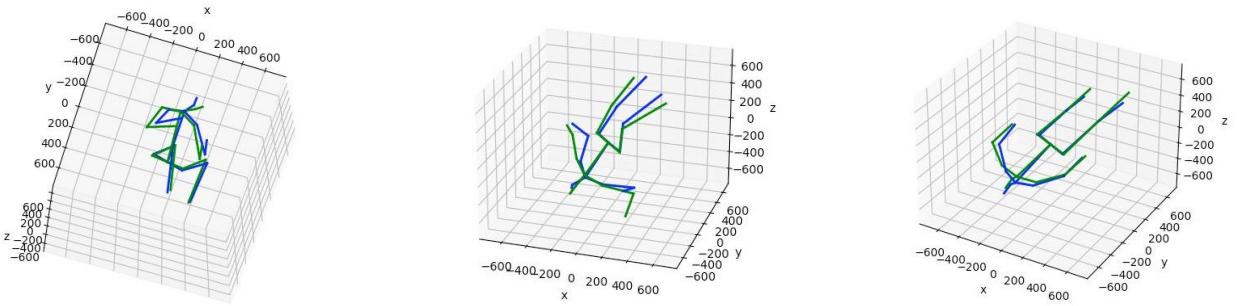

Figure 4: Samples from the validation set using approach (E). Green are the ground-truth joints locations and Blue are the predicted locations. We can clearly see that for the most right image, where we have a rather easy pose, the prediction performs better than for more complex poses such as the one on the left and in the middle.

3 EVALUATION

For the evaluation of the previously defined methods, we used the provided datasets Human3.6M and MPII. For Human3.6M, the training dataset has 40896 samples containing subjects 1, 5, 6, 7. For MPII, the training set consists of 22246 samples. For validation in order to prevent overfitting, we used the Human3.6M validation set containing 8140 samples of subject 8. As a last step for the final evaluation metric, the trained models are evaluated on the test set, containing 8608 RGB images from the Human3.6M dataset (subjects 9, 11), having some camera properties but no ground-truth labels available. For project submission, we predict the 3D coordinates of 17 body joints for each sample, resulting in a total of 51 coordinates for each of the 8608 samples from the test set. The evaluation metric for this project is the root-aligned mean per joint position error (MPJPE) in millimeter which is computed as follows: 1) first, the predicted coordinates are translated to the original image coordinates. 2) next, by using the camera intrinsics and the ground-truth root (pelvis) depth, the image coordinates are projected back to the camera space. 3) both the predicted key-points and the ground-truth labels are aligned relative to their root (hence subtracting the 3D root pelvis coordinates from all joint coordinates). 4) for each joint, the per-joint L2-norm is computed and the mean for each sample computed. Then, the global mean is computed over all samples:

$$MPJPE = \frac{1}{M}\sum_{i=1}^{M} \|p_i - \hat{p}_i\|$$

Where p_i is the ground-truth joint location, $\hat{p}_i$ is the predicted joint location in the root-aligned euclidean 3D space, and M is the total number of samples. The norm here is defined to be the L2-norm. The results for the test set of the Human3.6M dataset are reported in Table 1.

We take the provided boilerplate code (A) as a baseline and improved step by step upon it. Using data augmentation (B) by performing random flipping and rotation for the provided training dataset only helps to decrease the error by a fraction. Approach (C) implements the work 'Simple Baseline' by Martinez et al. [4] and takes ground-truth 2D labels for the training of a 3D regression. Approach (D) and (E) use a two-stage approach where the RGB image is first fed into a self-trained Stacked Hourglass (SH) model for a 2D prediction, which is then normalized and fed into a Simple Baseline (SB) model for the 3D regression. For (D), we take the full output of the SB method for the submission. For (E), we extract only the depth coordinate but keep the x and y-coordinates from the SH output (which was fed into the SB model for the 3D uplifting) before transforming back to the camera space. For (F) and (G), we evaluated the performance of the Integral Regression (IR) model where for (F) we took the full prediction from the IR model and for (G) we only extracted the x, and y-coordinate from the IR output, but used additionally the SH model together with the SB model for an uplifting and estimation of the depth coordinate. This method uses at the end three trained models, but doesn't use any ensemble techniques. (G) showed the best results so far for us, and we argue that it may be the most robust model from all the ones we experimented with. By further exploiting the properties of the Human3.6M dataset and considering its multi-view setting, where we have multiple RGB images from the same action taken at different angles, we were able to even further reduce the MPJPE by a large margin for (H). Unfortunately, this last submission is not considered as a valid submission within the context of this project and we noticed this too lately before the final deadline of project 2. Regardless the invalid submission, we thought it is worth to be mentioned that we could roughly reproduce the results from [6], reporting around 27mm for their MPJPE on the Human3.6M test dataset.

4 DISCUSSION AND CONCLUSION

Within this project, we were able to reproduce the results of multiple state-of-the-art methods from the past few years, namely for the Stacked Hourglass (SH), Simple Baseline (SB), and Cross-View method. It is interesting to see that method (E) — where we only extract the z-coordinate (i.e. the depth) form the 3D regression — showed better results than for (D), where we took the full coordinates from the regression output. We think this might be due to the MSE loss, which is minimized during learning, but introduces further noise for the final predictions. As the Simple Baseline method uses a simple L2-loss term for their learning, we tried to come up with a more restricted loss term making usage of certain bone length priors such as the one proposed in [7], which proposes

Table 1: Results on the Human3.6M test set as reported by the submission webpage for project 2, GT: Ground Truth, SB: Simple Baseline, SH: Stacked Hourglass, IR: Integral Regression

Method	MPJPE (mm)
(A) Baseline (x, y, z)	149.81
(B) Baseline with data augmentation (x, y, z)	132.02
(C) GT ⟶ SB (x, y, z)	71.68
(D) SH ⟶ SB (x, y, z)	72.54
(E) SH (x, y) ⟶ SB (z)	64.02
(F) IR (x, y, z)	63.26
(G) **IR (x, y) + SH ⟶ SB (z)**	**60.82**
(H) Cross-View (x, y, z)	37.65

a structure-aware regression approach considering the connection between certain joints and defining a relative loss between different joints. However, we were not able to reproduce the same low error as they reported.

Performing 3D Human pose regression solely from monocular RGB images is not a trivial task, especially under occlusion and when multiple people are in the scene, such as for in-the-wild RGB images. Making use on a prior for the connections of different joints makes intuitively sense and it would be interesting to see how such a constrained loss term would improve the performance for 3D Human Pose estimations in more complex environments. Further, with the advent of larger and more complex datasets and the wider adoption of RGB-D systems, it will be interesting to see how the estimation of 3D Human poses will improve in the upcoming years.

A FINAL SUBMISSION

For our final submission, we decided to go with the method with the second lowest submission score on the project page. Our lowest score marks clearly method (H), where we implemented a network exploiting the property of multi-view geometry, but does not mark a valid submission. The approach with the second lowest MPJPE uses a combination of three self-trained networks, namely a Stacked Hourglass network, a Simple Baseline network, and an Integral Regression network. This combination does not mark an ensemble, but acts as a multi-stage pipeline: First, we trained the Stacked Hourglass network on both Human3.6M and MPII the same way as described in [5] in terms of data augmentation and normalization of the input patch of size 256x256. This gives us already reasonable predictions for the x- and y-coordinates in pixel space. Next, we normalize that 2D output by subtracting the mean and dividing by the standard deviation independently for each of the 17*2=34 joint coordinates. Then, we train the Simple Baseline network with this data for an uplifting to 3D and denormalize the output of shape 17*3=51 joint coordinates with the statistics on the ground truth 3D data of Human3.6M. From this output, we only extract the z-coordinate of each of the 17 joints and concat it together with the x- and y-coordinates of the 3D prediction of the Integral Regression network. SH is trained on the concatenated training and validation datasets of both Human3.6M and MPII (71750 samples) and ran for 70 epochs. The network size was set to $N = 2$ and the training and validation batch size was set to 6 samples. The Learning Rate started at 5e-4 and was reduced after 20 and 30 epochs respectively with a factor of 0.1. SB is trained with an L1-Loss on the output of the previously trained SH model for the concatenated dataset of the training and validation set of Human3.6M. Batch size was set to 256 for the training phase (without a validation phase) for 192 epochs. Learning rate started with 1.0e-3 and was decayed by a factor 0.96 after 31200 steps. The Integral Regression model was trained on both MPII and Human3.6M training datasets with a starting learning rate of 0.001 and decay of 0.1 after epoch 270 and 290. Patch width was set to 256x256 and data augmentation was set according to [8]. We used an L1 loss for training. All three networks were trained with ADAM as an optimizer with the remaining parameters of ADAM for Pytorch (torch==1.8.1+cu111) left as default. SH and SB were trained on the Leonhard cluster, IR was trained on an NVIDIA Tesla V100 with 32GB GPU RAM available, rented from the company ExaMesh GmbH, as the Leonhard Cluster was nearly unavailable for us in the last weeks before the project deadline.

B ACKNOWLEDGEMENT

We would like to thank BOW IT&T GmbH for their support in providing compute resources with an Nvidia GeForce RTX 2070 during the time of this project.

REFERENCES

[1] "Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt" Schiele. 2014. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Hai Ci, Xiaoxuan Ma, Chunyu Wang, and Yizhou Wang. 2020. Locally connected network for monocular 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[3] 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.

[4] J. Martinez, R. Hossain, J. Romero, and J. J. Little. 2017. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 2659–2668. https://doi.org/10.1109/ICCV.2017.288

[5] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. arXiv:cs.CV/1603.06937

[6] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. 2019. Cross View Fusion for 3D Human Pose Estimation. arXiv:cs.CV/1909.01203

[7] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 2602–2611.

[8] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. 2018. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 529–545.

[9] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. 2019. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3425–3435.

[10] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*. 398–407.