3D Reconstruction of Humans from Images

Severin Husmann
shusmann@student.ethz.ch

Henry Trinh
trinhhe@student.ethz.ch

Daniel Végh
davegh@student.ethz.ch

ABSTRACT

In this project report, we describe the employed methods to tackle the task of creating human 3D reconstructions as meshes based on a skinned multi-person linear model (SMPL) [8]. We trained on a partition of Human3.6M [5] dataset provided for the project task. For the given data, we then implement further data augmentation techniques. To achieve our best result of a public test set score of 0.0197, we employ a GAN based architecture for generating more and more human like poses for the given 2D images. On this architecture we conduct hyper parameter tuning for the different losses and specifically downsampling to tackle temporal data set similarities as well. Other implemented methods to stop the model from overfitting are early stopping and learning reate scheduling. As future research directions, we suggest diving deeper into potential data augmentation techniques as well as other possibly more appropriate loss functions.

1 INTRODUCTION

Model-based human pose estimation is currently approached through two different paradigms. Optimization based methods fit a parametric body model to 2D observations in an iterative manner, leading to accurate image model alignments, but are often slow and sensitive to the initialization. In contrast, regression-based methods, that use a deep network to directly estimate the model parameters from pixels, tend to provide reasonable, but not pixel accurate, results while requiring huge amounts of supervision.

In this project, the goal was to conduct 3D reconstruction of human poses directly from 2D images by leveraging the skinned multi-person linear model (SMPL) [8]. In recent years, researchers explored this problem since generated 3D human poses has a wide range of possible applications such as movies and games to virtual and augmented reality. With the intricacies of the shape and pose of humans, the problem is highly complex. SMPL is a model that allows to project the human body model (i.e. its vertices) from its pose parameters and shape coefficients. Nonetheless, the complexity of the problem is still very high as many of the from model generated poses are unrealistic for humans.

In previous works such as SMPLify [2], this was tackled using three pose priors and a shape prior in an objective function minimization problem. Namely, in the pose priors they introduced a term that penalizes unnatural rotations, one that makes certain (natural) poses more likely, and finally one that penalizes self-penetrating poses conditioned both on pose and shape. Through the shape prior,the model forced is nudged to converge towards a mean shape. Furthermore, there was also work using a combination of parameter regression followed by such a mentioned an optimization procedure [7].

However, we believe that the model itself should learn what realistic human 3D meshes and poses are. For this reason, we mainly followed [6] which introduces a discriminator that implicitly learns pose priors and judges generated poses to be real or as created by the generator model. Hence, we used a stronger form of weak supervision approach to tackle this problem and introduce supervision on the realism of a generated pose with the discriminator. In their paper though, the authors used other motion capture datasets that allow them to train the adversarial prior extensively. With our constraint to only use the Human 3.6M dataset, we combat this issue in our end-to-end learning framework by data augmentation inspired by the SPIN paper[7].

2 METHOD

In this section we describe the various steps that led to our final results on the public test dataset. First, we discuss the applied data augmentations, followed by the neural network architecture description as well as implementation details (i.e. hyperparameters) with our reasoning on them.

2.1 Data augmentation

There are different 2D (e.g. LSP, LSP-extended, MPII, MS COCO) and 3D (e.g. Human3.6M, MPI-INF-3DHP) annoted datasets to train and evaluate 3D human body models. For our task we were given a partition (i.e. subjects S1, S5, S6, S7, S8) from the 3D Human3.6M dataset to work with. Due to pose similarities in the different subjects and thtions as well as to handle over-fitting during training we apply several data augmentation techniques both to the images and to the pose parameters (body/hand). First we deterministically re-sample the original dataset choosing different ratios (e.g. 0.2, 0.5, i.e. take every 5th or 2nd image, respectively) to deal with the highly similar poses generated in quick succession by the actors, then we load and crop images according to the ground truth bounding box with resolution 224x224. On this imagenet normalization is then applied. After that, we use random rotation, scaling and flip to the image and pose parameters. Moreover, we also add random pixel noise in channel-wise manner to the images. We also experiment by adding random gaussian noise and with normalizing pose/beta parameters to no success. The data augmentations technique was inspired by the SPIN github release [7].

2.2 Model

As a first step, we substitute the default image encoder layer (backbone) of the provided 3D mesh creation model with a more accurate 2D CNN pretrained model. We experiment with various backbone models like VGG, ResNet18, ResNet50, Inception etc. from the official Pytorch library combined with the default linear layer to infer the SMPL parameters. Out of all the neural networks, ResNet50 gave us the quickest convergence and best result.

Although we achieve better results with the improved image features, the learnt ϕ parameters constrains the space of possible solution and variations. Motivated by [3, 4, 9] we introduce a parameter regressor that iteratively infers SMPL θ and β parameters. The regressor takes as input the image features concatenated with

mean shape coefficients $\beta \in \mathbb{R}^{10}$ and mean poses $\theta \in \mathbb{R}^{24*3}$ from the neutral SMPL body model. As mentioned in [3, 4, 9] those parameters include rotations which makes it hard to regress the parameters in only one iteration. The parameter regressor consists of two fully-connected layers with 1024 neurons and ReLu activation functions each with a dropout layer in between. The last layer is a fully connected layer predicting our parameters. We try to use three iterations as they recommended but we did not get good results similar results. After trying a variety of iterations, we get the best result with five iterations.

For further improvement, we consider using different loss functions. So far the default vertex-to-vertex L1 and L2 loss functions seem to only supervise the body parameters implicitly. We observe that the current loss does not take into account the shapes and poses from the human. It means that the optimum can be reached with infeasible human configurations. To address this, we introduce more loss functions that correspond to the shape coefficients β, poses θ, 3D-positions of the joints and the onto the image projected 2D joint positions. Thus all our loss functions are:

$$L_{v2v_L2} = ||v_i - \widehat{v_i}||_2^2 \tag{1}$$

$$L_{v2v_L1} = ||v_i - \widehat{v_i}||_1 \tag{2}$$

$$L_{\beta} = ||\beta_i - \widehat{\beta_i}||_2^2 \tag{3}$$

$$L_{\theta} = ||R(\theta_i) - R(\widehat{\theta_i})||_2^2 \tag{4}$$

$$L_{3Djoints} = ||X_i - \widehat{X_i}||_2^2 \tag{5}$$

$$L_{2Djoints} = ||x_i - \widehat{x_i}||_1 \tag{6}$$

where $\widehat{vi}$ are ground truth mesh vertices, R is the Rodrigues transformation, X_i joints in 3D-coordinates and x_i joints in 2D-coordinates. Ground truths β and θ are given in the dataset. We get the ground truth vertices and 3D joints by forwarding β and θ into our SMPL model and apply the joint regressor matrix on the mesh vertices. 2D joints we retrieve by applying a function that projects the 3D joints into 2D joints with the camera focal lengths and its center. The final loss function is then the summation of all the above mentioned loss functions where each of them are weighted according to our needs.

Visualizing the mesh projection onto the images, we observed that the arms were rather thin or some poses were infeasible e.g. the torso completely turned by 180 degrees even when using all the mentioned loss functions. It seems as if having only a neural network with an iterative parameter regressor and the mentioned losses, is not enough to reconstruct accurate human shapes in some cases. One can observe that for human bodies some joints like shoulder and elbows exhibit higher degree of rotations. As a consequence linear blend skinning may generate unrealistic human body meshes given by the extreme joint rotations. To put some constraints on those joints, we take inspiration from [6] where they introduce a discriminator to set those constraints.

The discriminator trains separately on the pose and shape parameters. We have for each joint one discriminator that trains on one joint to learn its angle limits and also a discriminator that trains on all joints together to get the joint distribution of the entire kinematic tree. Their approach makes it possible that we do not have to make any assumptions about the humanly possible joint limits since it is learned during training. As it is very common in GAN for mode collapse, their approach eliminates it since the network has to fool both discriminator and also minimize the $L_{2Djoints}$. Overall, we thus transform our model to a GAN model where the generator G is the backbone neural network with the parameter regressor and 25 discriminators (23 joints, 1 for shapes, 1 for all joints). The generator has then following adversarial loss function:

$$min\, L_{adv}(G) = \sum_i \mathbb{E}_{\beta,\theta\sim p_G}[(D_i(G(I)) - 1)^2] \tag{7}$$

and each discriminator tries to decrease:

$$min\, L(D_i) = \mathbb{E}_{\beta,\theta\sim p_{data}}[(D_i(\beta,\theta) - 1)^2] + \mathbb{E}_{\beta,\theta\sim p_G}[D_i(G(I))^2] \tag{8}$$

In the implementation the discriminator for the shape β are two fully-connected layers with 10, 5, and 1 neurons. For the pose θ discriminators we convert with the Rodrigues function all our axis angles in $3x3$ rotation matrices and feed it into two fully-connected layers with 32 hidden neurons. Then we forward it to 23 different discriminators that output 1D values. The discriminator for all the whole pose distribution concatenates all 23 * 32 internal outputs of the previous discriminators through another two fully-connected layers of 1024 neurons and outputs the final 1D value.

In summary our model consists of two parts, namely a generator which takes as input an image and generates the SMPL parameters (shape coefficients $\beta \in \mathbb{R}^{10}$ and poses $\theta \in \mathbb{R}^{24*3}$ in axis-angle representation) and a discriminator telling us if those parameters are giving us human-like meshes.

3 EVALUATION

Due to the extent of dataset and time required to run each of the experiments on the full dataset, we trained different models in different settings reducing training time, while effect of hyperparameter tuning we still kept comparable. Therefore, it is hard to find a common base for comparison of each experiments to the end. Still we summarize our results in general, based on the validation L2 score over the runs with different configuration. The experiments gave the best results for the SMPL model with GAN regressor achieving 0.023, and 0.020 test score in case of 0.5 re-sampled dataset respectively (see summary in Table 1). This also demonstrates the effect of our model overfitting due to more training on the validation L2 loss.

4 DISCUSSION

In this section we discuss our results and we conclude by mentioning further directions.

As can be seen from Table 1 our best model achieves a public test score performance of 0.020. Before that, we continuously tried to improve upon our validation loss and employed different combinations of models and hyperparameters to achieve the lowest possible validation score without overfitting. With such an extensive dataset, however, it is difficult to do extensive experiments. For this reason we were guided by [6] for the GAN parameter regressor and settled with a slightly different configuration due to an observed improvement. We also tried out other settings for instance for downsampling the dataset, in the end taking the best

Model	Optimizer	Backbone	Data augmentation		Resample	Val L2 loss	Test score
			Image	Pose			
SMPL	SGD	default	na	na	na	0.251	0.031
SMPL	SGD	resnet18	true	false	na	0.056	0.077
SMPL	Adam	resnet50	true	true	na	0.041	0.052
SMPL with iterative regressor	Adam	resnet50	true	true	na	0.023	0.035
SMPL with GAN regressor	Adam	resnet50	true	true	0.5	0.009	0.023
SMPL with GAN regressor	Adam	resnet50	true	true	0.5	0.011	0.020

Table 1: Table showing L2 validation scores, and final test score for different models and configuration

observed combination. Nonetheless, a step to further improve our performance could be a more thorough hyperparameter search.

We also observed that due to the lack of appropriate data, the model can start to overfit at a certain L2 validation score. Hence, we lowerbounded the training for our full training script in order to get around this issue.

Future research. Since the Human 3.6M dataset is constrained in terms of poses and corresponding images (i.e. the same pose has many similar images), we think that further data augmentation could provide to be very helpful. Our data augmentation pipeline together we deterministic downsampling definitely helped in tackling this issue as the results for our method were worse on the full dataset and without data augmentation. Nonetheless, there are various directions which could be explored for further data augmentation. An excellent approach could be to follow the method in [10]. In this paper, the author's only use 3D meshes and sample random camera parameters to project the 3D pose onto an image plane to retrieve 2D ground truth data. This dataset could for instance be used to pretrain the discriminator in our architecture. Other options for image data augmentation could be changing the backgrounds from the images to have them in more diverse habitats.

Furthermore, an in-depth exploration of more advanced loss functions could also lead to better results. We mainly explored rather simple supervised losses from the predicted data directly (i.e. L1 and L2 terms for the different predictions). One option that we think could possibly be used would be the 3D Chamfer distance as loss function instead of a the vertex-to-vertex losses.

Finally, we would also explore the structured prediction layer introduced by [1] as an additional layer in our model as the paper showed to improve upon all back then state-of-the-art results and we believe such more structured prediction approach could further improve our iterative parameter regressor that does not yet take into account joint relations.

5 CONCLUSION

Overall, the goal of this project was to explore deep learning techniques for 3D reconstruction of humans from images. For this reason, we implemented a generator-discriminator architecture to force the generator towards more humanly feasible poses and natural shapes. While overall our results with a best public test set vertex-to-vertex L2 score of 0.01967 are very positive, there is still room for improvement. In this regard, we suggest more in-depth researching data augmentation techniques, further loss function possibilities as well as applying structured predictions on the pose parameters.

REFERENCES

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. 2019. Structured Prediction Helps 3D Human Motion Modelling. *CoRR* abs/1910.09070 (2019). arXiv:1910.09070 http://arxiv.org/abs/1910.09070

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. *CoRR* abs/1607.08128 (2016). arXiv:1607.08128 http://arxiv.org/abs/1607.08128

[3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. 2016. Human Pose Estimation with Iterative Error Feedback. arXiv:cs.CV/1507.06550

[4] Piotr Dollár, Peter Welinder, and Pietro Perona. 2010. Cascaded pose regression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1078–1085. https://doi.org/10.1109/CVPR.2010.5540094

[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] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. (2018). arXiv:cs.CV/1712.06584

[7] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. *CoRR* abs/1909.12828 (2019). arXiv:1909.12828 http://arxiv.org/abs/1909.12828

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[9] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. 2015. Training a Feedback Loop for Hand Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[10] Jie Song, Xu Chen, and Otmar Hilliges. 2020. Human Body Model Fitting by Learned Gradient Descent. *CoRR* abs/2008.08474 (2020). arXiv:2008.08474 https://arxiv.org/abs/2008.08474