Machine Perception Report

Project 3: 3D reconstruction of humans from images

Alessandro Maissen
maisseal@student.ethz.ch

Mengdi Wang
menwang@student.ethz.ch

Aygul Zagidullina
azagidull@student.ethz.ch

ABSTRACT

In this project, we approach 3D reconstruction of humans' body shape in natural poses from 2D images by leveraging convolutional encoder, iterative regressor and SMPL model sequentially. Central part of our work is the incorporation of the convolutional encoder (ResNet-50 or ResNet-152, [5]) within the end-to-end framework, and the customization of an iterative regressor [8]. The proposed approach outperforms the simple baseline model on this task and offers an attractive solution for generating 3D human meshes.

1 INTRODUCTION

The reconstruction of realistic full body 3D human pose and shape is a crucial task within the computer vision domain, given an increasing range of applications nowadays (e.g., human-robot interaction, gaming, virtual reality). However, the inherent ambiguity of the 3D information poses a challenge for recovery of an accurate human body geometry from a single input image. Similar image projections can be derived from completely different 3D poses due to the loss of 3D information. Moreover, multiple variations of the image, e.g., different environments and viewpoints yield multiple solutions. Thus, due to the particularly challenging nature of such a problem, the literature remains sparse. There are two main branches of research that aim at reconstruction of 3D human bodies, the optimization-based methods, and the learning-based methods. In turn, the learning-based methods either rely on the 3D human models, such as the SMPL model (i.e. model-based methods), or directly regress the 3D human body representation from an image (i.e. model-free methods). The model-based learning methods are computationally efficient compared to the model-free learning approaches and optimization-based techniques. Furthermore, these methods can make use of the prior knowledge embedded in the 3D human model, and hence, reconstruct biologically more plausible human bodies compared to the model-free techniques.

In our project, we consider the SMPL model-based learning method. This approach provides the above mentioned efficiency and plausibility advantages. Our key contribution is the implementation of the transfer learning (by e.g. usage of the pre-trained image-based convolutional models) for the feature extraction. This feature extraction technique ensures high accuracy in the estimation of input parameters for the SMPL model.

2 RELATED WORK

For completeness, we briefly discuss here the above-mentioned related work with respective references.

2.1 Skinned Multi-Person Linear Model

SMPL (Skinned Multi-Person Linear Model (SMPL) [11]) is a statistical body model for generating 3D human meshes. The parameters of the SMPL model are the human pose (per-joint 3 degrees of freedom, axis-angle representation) and human shape (10 degrees of freedom, principal components) coefficients. The objective of the project is to develop a Neural Network architecture that predicts these input parameters with high accuracy in order to retrieve the ground truth human mesh.

2.2 Optimization-based Methods

Originally, the 3D human body reconstruction was solved by optimizing parameters of a predefined 3D human mesh models (e.g. SMPL) with respect to the ground-truth body landmark locations [4], or employing a network for 2D keypoints estimation [1].

Nowadays, most approaches also rely on iterative optimization that attempts to estimate a full body 3D shape that is consistent with 2D image observations, despite the significant run-time required for optimization and common failures because of local minima.

2.3 Model-based Methods

The model-based approaches directly incorporate a parametric 3D human model (e.g. SMPL) into the architecture. This way, the problem at hand is reduced to model

parameter estimation. For example, within the Human Mesh Recovery (HMR) [8] framework the SMPL parameters are directly regressed from an image. Recently, SPIN (SMPL oPtimization IN the loop) [10] incorporates 3D human model parameter optimization into network training process by supervising network with respect to the optimization result. Hence, SPIN is the combination of the model-based and optimization-based approaches. This latter architecture achieves the state-of-the-art results among model-based 3D human body estimation approaches.

Compared with optimization-based methods, these architectures are more computationally efficient, and can reconstruct more human-alike body meshes. However, the representation capability of these neural networks is in a sense limited by the parameter space of the predefined human models.

2.4 Model-free Method

The model-free methods do not rely on human models and directly regress 3D human body representation from an image. The most recent volumetric (BodyNet, [13]) and point cloud based representations [3] are flexible and can retrieve 3D objects with different topological structure. However, the detailed surface reconstruction is limited by storage, and more importantly, the human mesh vs. image correspondence might be poor.

3 DATA

The data we use is a subset taken from Human3.6M [7]. The dataset consists of images of 7 subjects. Subjects 9 and 11 are used as a test set and remaining subjects (1, 5, 6, 7, 8) as a training set. For the model selection purposes we, furthermore, use subjects 7 and 8 as a validation objects. As part of the project, there has not been any preprocessing or change of representation performed.

4 METHOD

The following steps represent the end-to-end proposed architecture in detail:

step 1 feature extraction via transfer learning: use the pre-trained convolutional encoder (e.g. ResNet-50 or ResNet-152) with parameters not being updated and the last network layer replaced by an identity layer;

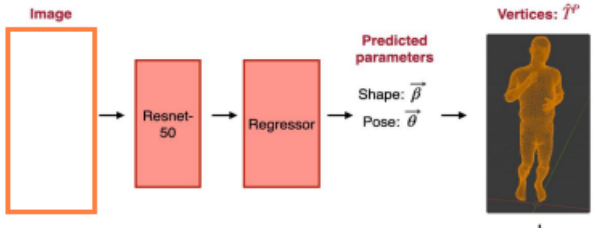

Figure 1: Overview of the basic end-to-end framework. A single image is used as the input to the convolutional encoder (e.g. ResNet-50 or ResNet-152). The regressor layers transfer the output of the encoder into the predicted shape β, pose θ and other parameters. These parameters are subsequently used to reconstruct vertices by the SMPL model. Based on figure by [6].

step 2 prediction of the SMPL model parameters (pose, shape and orientation) via linear regression layers: use the convolutional encoder output, $Y \in \mathbb{R}^{2048}$, as an input feature vector to the linear layers. The linear regressors are either stacked sequentially, i.e. $2048 \rightarrow 1024 \rightarrow 1024 \rightarrow 82$, or represent the iterative module as in the HMR [8] setup;

step 3 generation of the 3D human body mesh geometry: use the output of the last layer, $Z \in \mathbb{R}^{82}$, as the parameters of the SMPL model in order to produce the $N = 6890$ vertices and compare with the true mesh vertices.

Finally, above 3 steps define our model, which is trained with respect to the L_2 or/and L_1 loss.

The schematic representation of the described architecture is given in Figure 1.

The iterative regressor, unlike in the original HMR framework, is based on the minimization of the loss w.r.t. the mesh vertices and not the joint re-projection error since the 3D joint location annotations are not available in our setup.

We start from an initial estimate of the body parameters $\Psi_0 = \bar{\Psi}$, where $\bar{\Psi}$ represents the mean [*], which is concatenated to the extracted features Y from the convolutional encoder and feed this vector to a neural network that predicts an update $\Delta\Psi_1 = NN([Y, \Psi_0])$. The new parameter value is then equal to $\Psi_1 = \Psi_0 + \Delta\Psi_1$, and the whole process is repeated. As in [8], we iterate for 3 times. The overview of the iterative regressor approach is provided in Figure 2.

[*] We compute the mean based on the training objects' available parameters.

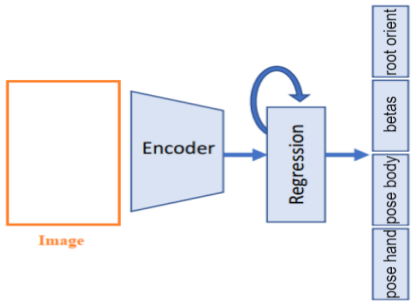

Figure 2: Overview of the enhanced iterative regressor framework. A single image is used as the input to the convolutional encoder. The output of the encoder is sent to the iterative 3D regression module that infers the latent 3D representation of human body that minimizes the L_2 or L_1 loss of the predicted mesh vertices w.r.t the true mesh vertices. Based on figure by [8].

In addition, given that the true 3D mesh parameters are also available, we tried to construct separate loss functions as per body shape, body pose, hand pose and body location, and based on that to produce the total weighted loss function. The idea is inspired by the Ex-Pose [2] combination "of a 2D re-projection loss, 3D joint errors and a loss on the parameters". However, we do not use the 2D re-projection loss.

4.1 Implementation Details

According to the architecture, the following hyperparameter values are considered during implementation and model tuning: learning rate, dropout rate in the iterative regressor module, batch size, number of regressor iterations and number of epochs.

We train our networks using two different optimizers: Adam [9] or SGD. The entire pipeline described above is implemented in PyTorch [12] and by default ported onto GPU.

5 EVALUATION

In Table 1 we present the 3 architectures that yield the top 3 validation and public scores along with the hard and easy baselines.

The models at this stage are trained on a subset containing three subjects. The left two subjects are used for the validation, and hence, model selection.

As expected, the ResNet-152 is more powerful in feature extraction than the ResNet-50, and the iterative regressor module improves performance compared to the sequential linear layers.

Architecture	Validation Error	Public Score
ResNet-50 + sequential regressor	0.0192	0.0243
ResNet-50 + iterative regressor + weighted loss	0.1998	0.0257
ResNet-152 + iterative regressor	0.0169	0.0221
Hard baseline	–	0.0327
Easy baseline	–	0.0388

Table 1: Evaluation on the subset of Human3.6M dataset. The numbers are the standard mean squared error between predicted mesh vertices and the ground truth mesh vertices.

Hence, according to the table above, the best scores are provided by the ResNet-152 encoder combined with iterative regressor module. Therefore this neural network was chosen for the final submission.

6 FINAL SUBMISSION

Our final submission is based on the following architecture and hyperparameter choices:

- convolutional encoder: ResNet-152
- iterative regressor layers: 3
- optimizer: Adam
- learning rate: 0.0001
- batch size: 64
- number of epochs: 10
- dropout rate: 0.50

During the model selection process, based on the validation scores, we observe that the model does not overfit until epoch 10 with the above hyperparamters. For the final submission, we trained the model on all the (1, 5, 6, 7, 8) objects. The final submission score is 0.02094.

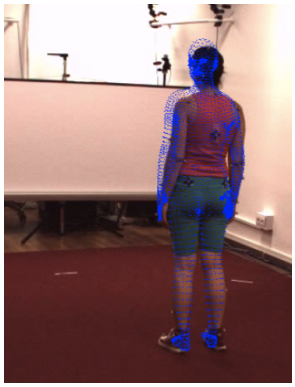

Figure 3: 3D human body mesh reconstruction on a validation object obtained from ResNet-52 + iterative regressor + parameter-based loss architecture.

7 DISCUSSION

As described earlier, we take the weighted SMPL parameters based loss also into account during training. In such case, although the reconstructed 3D human mesh looks very human alike and potentially is more plausible then the outcome of the other models under consideration (see Figure 3), the validation scores obtained with this architecture are worse compared to the model chosen as a final submission. The reason for that may lie in the lack of weights tuning for the combined loss given the limited computational resources according to our assumption. Potentially, one could construct a grid of weight values and optimize the weighted loss and training process with respect to the weight parameters.

8 CONCLUSION

In our project, we consider the SMPL model-based learning method for the 3D body mesh reconstruction. We implement the transfer learning for the feature extraction and the iterative regressor for the SMPL model parameters estimation. This network architecture yields the MSE of 0.02094 on the public leader board.

REFERENCES

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. arXiv:cs.CV/1607.08128

[2] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2020. Monocular Expressive Body Regression through Body-Driven Attention. arXiv:cs.CV/2008.09062

[3] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. 2019. Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images. arXiv:cs.CV/1908.00439

[4] P. Guan, A. Weiss, A. Balan, and M. J. Black. 2009. Estimating human shape and pose from a single image. In *Int. Conf. on Computer Vision, ICCV*. 1381–1388.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[6] Qi He. 2020. *Three-dimensional reconstruction of human body via machine learning*. Ph.D. Dissertation.

[7] 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.

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

[9] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:cs.LG/1412.6980

[10] 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. arXiv:cs.CV/1909.12828

[11] 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.

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[13] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. BodyNet: Volumetric Inference of 3D Human Body Shapes. arXiv:cs.CV/1804.04875