A Hybrid Method for 3D Human Pose Estimation

Maximilian Pittner
max.pittner@tum.de

Valentin Starlinger
valentin.starlinger@tum.de

Abstract

We describe a Hybrid Method for reconstructing a full 3D human body represented by a mesh from a single RGB image. The 3D model is defined by parameters including information about joint poses and shapes. We predict these as well as the camera parameters by feeding the input image into a neural network consisting of a pretrained CNN and a fully connected network. For training our network, we reproject the predicted 3D mesh to the image plane and minimize a combined reprojection loss that compares the projected 2D joint locations and silhouette from our prediction to ground truth joint annotations and segmentation data. Additionally, we use a critic network to obtain feasible human body configurations.

1. Introduction

Estimating 3D human poses from single RGB images is a field of computer vision that has proven to be a challenging task. In recent years, researchers started to tackle this problem by using neural networks. A descriptive yet complex way of representing a human body is to use a mesh. There has already been quite success in predicting 3D meshes from single images, for instance by [1]. We propose to further advance these methods by combining different already existing ideas in order to improve the accuracy of these approaches.

2. Architecture

Our proposed hybrid network architecture is illustrated in Figure 1. We use the term *hybrid* as we combine different building blocks, mainly from [1], [2] and [3]. The main pipeline for the parameter estimation is adapted from [1]. In this part, the input image is fed into an encoder network for feature extraction, more precisely a ResNet50 [4] pre-trained on ImageNet [5] is used. The obtained features are then used as inputs for a regression network consisting of three fully connected layers to predict the parameters describing the 3D human mesh model. For the human body the SMPL model [6] is used which describes the mesh by 23 joint poses - thus $\vec{\theta} \in \mathbb{R}^{3\cdot 23}$ - and 10 shape parameters $\vec{\beta} \in \mathbb{R}^{10}$. Additionally, the six dimensional camera pose $\vec{\gamma} \in \mathbb{R}^6$ is estimated, so the network predicts 85 parameters $\vec{\Theta} = \{\vec{\theta}, \vec{\beta}, \vec{\gamma}\} \in \mathbb{R}^{85}$ in total.

2.1. Reprojection loss

For training, the 3D mesh is reprojected back to the image plane using the camera parameters and compared to the input image. Our reprojection loss involves a keypoint reprojection loss L_{kpr} as in [1], which calculates the L1 loss over projected joints and ground truth joints (keypoints); And a mesh reprojection loss adapted from [2], for which the segmentation data has to be given as ground truth. The mesh repojection loss L_{mr} computes the bidirectional distance between pixels of the ground truth silhouette and pixels of the silhouette resulting from the reprojection of the predicted mesh. More precisely, the loss is given by

$$L_{mr}(\vec{\theta},\, \vec{\beta},\, \vec{\gamma};\, S_{GT}) = \sum_{\vec{x} \in \hat{S}(\vec{\theta},\vec{\beta},\vec{\gamma})} d(\vec{x}, S_{GT})^2 + \sum_{\vec{x} \in S_{GT}} d\big(\vec{x}, \hat{S}(\vec{\theta}, \vec{\beta}, \vec{\gamma})\big)^2, \tag{1}$$

where $d(\vec{x}, S_{GT})$ denotes the absolute distance from a point $\vec{x}$ to the closest point belonging to the ground truth silhouette S_{GT} and $d\big(\vec{x}, \hat{S}(\vec{\theta}, \vec{\beta}, \vec{\gamma})\big)$ from point $\vec{x}$ to the closest point of the predicted silhouette $\hat{S}(\vec{\theta}, \vec{\beta}, \vec{\gamma})$.

The combined reprojection loss that we use is then given by

$$L_{comb} = w_{kpr} \cdot L_{kpr} + w_{mr} \cdot L_{mr}. \tag{2}$$

Our motivation for using this combined loss is justified in Section 4.1.

2.2. Critic Network

Without using some sort of discriminator, minimizing the reprojection loss would cause an enforcement of joint and shape fitting. The resulting predictions would not correspond to feasible body configurations as we show in Section 4.1. To prevent this we use a critic network based on ideas from [1] and [3]. In [1] a discriminator is used that takes in joint orientations and shapes. In order to make the generator learn feasible bone angles and lengths, we combine the discriminator from [1] with the critic network from

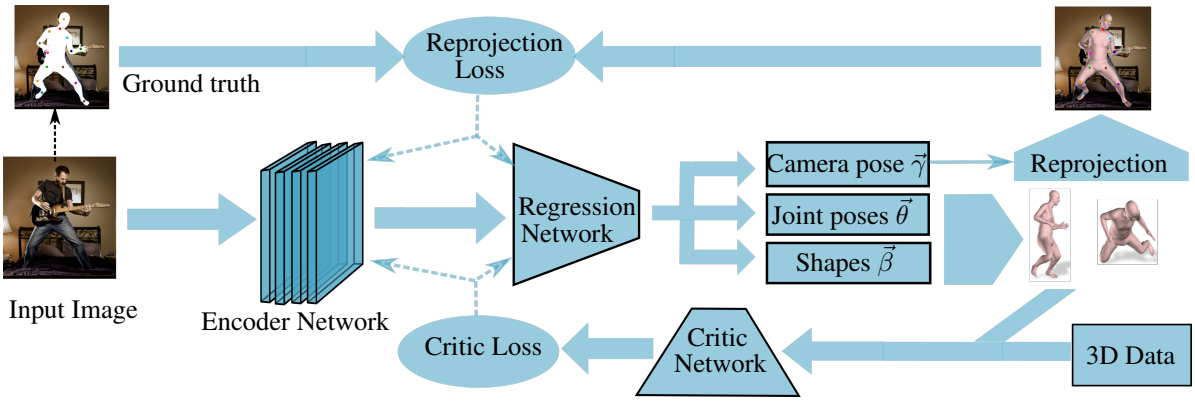

Figure 1. Our Hybrid network architecture.

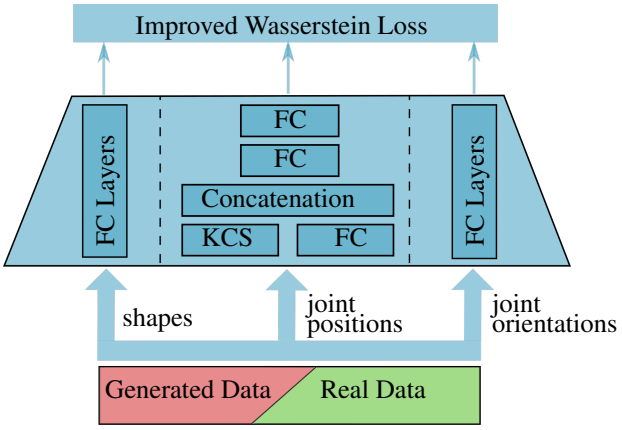

Figure 2. Our critic network containing building blocks of [1] (left and right stream) and [3] (middle stream). The joint positions and orientations can both be derived from $\vec{\theta}$.

[3] that uses a kinematic chain space layer (KCS) providing the information which joints connect to each other. Thus, the generator "knows" how a human skeleton is configured a priori and does not have to learn this constraint. Moreover, the critic network used in [3] uses the promising improved Wasserstein loss [7] to stabilize the training. We combine and slightly modify the networks from [1] and [3] and use the improved Wasserstein loss for training. Our resulting critic network is shown in Figure 2.

3. Datasets

The datasets used in this project were the Leeds Sport Poses (LSP) dataset [8] with 1000 train samples and 1000 validation samples and the Leeds Sport Poses extended (LSP extended) [9] dataset with an additional 8642 training samples for 2D images and corresponding joint keypoints as well as the segmentation data for these datasets provided by the UniteThePeople dataset [2]. Figure 3 shows an example of an image with the corresponding ground truth data for keypoints and segmentation.

The dataset used for training the critic network is the

Figure 3. An example image and ground truth from LSP/UniteThePeople dataset on the left side and one from the MoSh dataset on the right side.

MoSh dataset provided by [10] which contains approximately 4,000,000 training samples. They provided parameters of SMPL models for many different valid human poses. An example is illustrated in Figure 3.

4. Evaluation

4.1. Analysis of different reprojection techniques and critic network

To evaluate the mesh reprojection loss and the effect it can have, the network was overfitted on a small batch of 8 images under the following settings: Keypoint reprojection loss only (KPR only), mesh reprojection loss only (MR only) and a combined loss with both (combined).

In Figure 4, visual results for the different settings after overfitting are shown. Since the keypoint loss tries to minimize the distance between reprojected keypoints, this helps to get the limbs in the correct position but fails to preserve shape. The mesh reprojection loss tries to fill up the silhouette which prevents skinny mesh predictions but fails to get limbs into the right position. Finally, the combined loss tackles the issue of skinny mesh predictions as well as wrong body configurations.

Also shown in Figure 4 are results from overfitting using the above setting together with the critic network described in section 2.2. Without the critic network no feasible human body predictions are obtained. A critic network incorporating information about bone lengths and shapes gives almost feasible estimates. However, without considering feasible

Figure 4. From left to right, overfitting with KPR only, MR only, combined, critic without rotation, critic with rotation.

	Without Critic		
	KPR only	MR only	combined
KPR Loss	4.19	13.95	3.92
MR Loss	51.66	16.05	37.53
	With Critic		
	KPR only	MR only	combined
KPR Loss	4.70	18.25	4.41
MR Loss	50.32	92.44	44.03

Table 1. Validation losses for KPR only, MR only and combined loss with and without using the critic network.

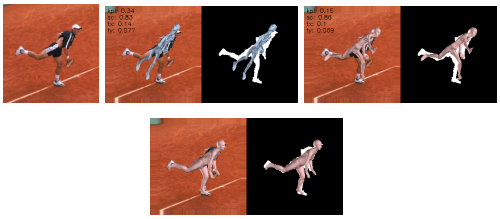

Figure 5. Visual comparison of KPR only (top middle), combined (top right) and baseline (bottom) for the input image illustrated top left.

	With Critic		Baseline[1]
	KPR only	Combined	
KPR Loss	3.85	3.20	2.64
MR Loss	33.65	30.02	27.09

Table 2. Validation losses after training for 120 epochs on LSP and LSP extended.

joint orientations the predictions end up with kinked hands and feet. Using a critic network considering bone lengths, shapes and joint orientations the generator is able to learn feasible human bodies.

4.2. Detailed evaluation of different approaches on LSP dataset

For a more detailed evaluation of the different approaches, the network was trained with the settings mentioned in Section 4.1 on the LSP dataset for 120 epochs. To measure the success of the different settings, for each one the validation score of the mesh reprojection loss and the keypoint reprojection loss are compared.

The results are shown in Table 1 and the patterns observed visually for the overfitting example are represented there as well. The combined loss without the critic network notably performed better on both of the measures compared to the configurations with the critic network. However, the visual results are much poorer and the good results are probably related to the critic not enforcing additional constraints on the predicted mesh. When comparing the results achieved with the critic network, both KPR only and combined loss performed very similar, suggesting that when using the critic network, the mesh reprojection loss has a much smaller effect.

4.3. Evaluation on LSP extended dataset

Since both the KPR only and combined loss performed well but similiar when training on the LSP dataset for 120 epochs, both configurations were trained on the combined LSP and LSP extended datasets. Additionally, the validation scores of the method presented by [1] were calculated as a baseline for both measures by using a pre-trained model. It is important to note that the provided model was trained for 75 epochs on a much larger dataset (which includes the LSP and LSP extended sets) of approximately 550,000 2D images and around 5,600,000 3D data samples.

With more data, and longer training the losses for our network went down for both combinations which is shown in Table 2. Similarly to the case above, the combined loss performed better than the KPR loss only approach. This suggests that the MR loss helps to improve the training of the network.

The baseline method still outperforms our approach, notably on the visual results as it does not show any twisted or squeezed humans as also shown in Figure 5 - especially when they are faced away from the camera or sideways. However, as discussed above, the baseline network was trained on a much larger dataset and with similar training our approach might also mitigate those issues.

5. Conclusion

This project tried to improve a baseline method for extracting 3D human meshes from 2D images by adding a mesh reprojection loss and changing the used discriminator to a critic network that additionally considers the skeleton configuration and provides a more stable training using the improved WGAN loss.

The evaluation suggests that with the current implementation the mesh reprojection loss does help to increase performance, but the baseline method still performed better. By training on a comparable amount of data our approach might outperform the baseline method which is suggested by the superior performance compared to the KPR only approach. Furthermore, using a scheduling approach for adapting the weights for the different losses during training might further increase the impact of the mesh reprojection loss and could be investigated in future work.

References

[1] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[2] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017.

[3] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. *arXiv preprint arXiv:1902.09868*, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[9] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[10] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, November 2014.