Machine Perception Report Marineprotection

Agisilaos Politis Leonardo Barberi Guglielmo Bonifazi Giovanni Acampa

ABSTRACT

This project aims to predict 3D SMPL pose and shape parameters of humans from RGB images. We use 2 datasets, 3DPW and MPII, containing that 3D and 2D annotations respectively. To solve the task of predicting 3D body parts, we took inspiration from the PARE paper, modifying the backbone, tuning hyperparameters and adapting their architecture to our training datasets [4]. After intensive training, our results performed better than the baseline.

1 INTRODUCTION

In the realm of computer vision and graphics, accurately estimating the 3D pose and shape of human bodies is a fundamental challenge. It holds immense significance in applications ranging from virtual reality and gaming to motion capture and human-computer interaction. To tackle this complex task, researchers have developed the SMPL model [5]. The 3D SMPL pose and shape estimation approach leverages a data-driven model to estimate the 3D pose and shape of a human body given a 2D image or video input. Despite substantial advancements, cutting-edge 3D human posture and form estimate techniques are still prone to partial occlusion and can make alarmingly inaccurate predictions, even when the majority of the body is visible. The PARE model was developed to solve the partial occlusion problem in 3D human pose and shape estimation, and we decided to adapt its architecture for this project [4]. Compared to other state-of-the-art 3D body model creation models, PARE is more robust to occlusions, making it a suitable fit for inspiration for the task at hand.

Our model uses a pre-trained backbone, that is a variation of ResNet50 and fine-tunes weights with data from the 3DPW and MPII datasets [6] [1]. After the backbone, the features are split into a 2D branch and a 3D branch using convolutional layers. These were used to apply a part-attention mechanism, combining the features again to finally predict the body pose and shape using one fully connected MLP per joint to predict.

2 METHOD

The ResNet-50 pre-trained neural network, available in the PyTorch library, was implemented as backbone and tested with different pre-trained weights [3]. One approach, was to use the weights of the DINO version of ResNet-50 [2]. DINO, whose stands for Self-**di**stillation with **no** labels, creates a teacher and a student network, both with the same architecture. It is very flexible and can be used with a Vision Transformers (ViT) or a ConvNet, such as the popular ResNet-50, and it is nowadays a popular choice as backbone. It can be found at https://dl.fbaipublicfiles.com/dino/dino_resnet50_pretrain/dino_resnet50_pretrain.pth. Our model was trained for a limited number of epochs trying both the DINO and the ImageNet weights. While theoretically promising, training with DINO weights resulted in significantly lower validation scores on our training data, hence the ImageNet weights were deemed a more suitable choice.

After this step, volumetric features were extracted by cutting the ResNet training before the global average pooling layer. As mentioned in the introduction, the features were then divided into a 2D and a 3D branch. For both the 2D and 3D branches, we used three $2\times$ upsampling followed by 3×3 convolutional layers, applied with batch-norm and ReLU. To obtain part attention maps, we applied $J + 1$ 1×1 convolutional kernels to 2D part features to reduce the channel dimension, where J is the number of joints to predict (that is 24). Each channel in the 2D output then represents one joint, and, upon applying a channel-wise softmax layer to the 2D features, each of these channels could be considered as an attention matrix, giving weights to pixels that are most important for predicting specific joints. The 2D features matrix was multiplied with the 3D features matrix, acting as a part-attention on the image features that are extracted from the 3D branch.

Finally, using 24 MLP for the poses, 1 for the shape and 1 for the camera parameters, we predicted the SMPL human poses and shapes. We used Adam optimizer with learning rate of 3×10^{-4} and batch size of 64. The model was trained for 120 epochs.

3 EVALUATION

We started to work studying some papers that already have good results on this task. Our first naive try was to predict the results only using ResNet-50 for feature extraction and some MLPs for pose, shape, and camera parameter regression. Of course the results were not satisfying, since the model's architecture was very naive to tackle the complex nature of the problem. We decided to follow the PARE model, adding the 2D, the 3D branches and the part attention, with a shallower network respect to PARE, using only two convolution blocks. Following the tips for training from the course, we also tried to implement stochastic weight averaging as a optimizer. Finally, we made the network deeper, we increased the learning rate and we trained the model for longer, also relying on tensorboard to verify whether the training was rolled out smoothly.

4 DISCUSSION

All the results are related to the MPJPE metric. The first result that was obtained during the experiments was a score of 112.74 for the ResNet + MLPs. After implementing the 2D and 3D branches and training only for 4 epochs, the score improved to 79.19, a significant improvement. After modifying the parameters and the backbone, making the model deeper and training for a higher number of epochs, we achieved a final score of 66.32, thus improving on the existing baseline.

5 CONCLUSION

In conclusion, our application and adaptation of the part-attention mechanism that was introduced in the PARE paper resulted in satisfying results, with much room to delve deeper into future investigation.

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] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

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

[4] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. 2021. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11127–11137.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

[6] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*.