Machine Perception Report MPGO

Mengtao Zhang Tristan Girand Maximum Wilder-Smith

ABSTRACT

In this report we present our model for 3D Human Pose estimation. It is a light-weight model based on deep convolutional neural network for image feature extraction. We use a resnet34 as backbone, then extract 3D features, 2D keypoint features and camera-shape features separately, using 2D keypoint features for supervision, then predict the pose with a HMR-style regressor.

1 INTRODUCTION

3D Human Pose estimation is an ill-posed problem which aims to directly regress 3D human features from the given RGB images. It is an active topic in computer vision and requires a lot of work to overcome issues such as self-occlusion, lack of depth information and so on. In the past few years, certain approaches have addressed these problems, such as the PARE[1] and HybrIK[2] models. The former model successfully solved the self-occlusion problem with partial self-attention mechanism to give weights to image features and regress based on that. The latter one produced more realistic human poses by using certain domain knowledge such as Inverse Kinematics. Instead of directly regressing the rotations, they regress the vectors representing 2D human poses and infer the rotation parameter using a previously defined Inverse Kinematics model. Both models have accomplished excellent performance and are seen as the state-of-the-art.

While reading the codes from those two models, we realise that both models have very complicated architecture and complicated losses to make the things work. In the contrast, our aim is using a simple and light-weight model which only based on convolutional neural network to get somewhat comparable results to those two state-of-the-arts. We still believe that CNNs are by far the most convenient and fastest way to deal with computer vision problems. So we decided to drop the complicated parts and tried to solve this task with pure CNN approach.

2 METHOD

The whole architecture is shown in Figure 1. The idea goes as follows: the shape and camera features are the relatively easy part and can be solved separately with some simple feature extractors and MLPs, leading separate branches for them. On the other hand, estimating the joint rotations is hard and requires a lot of work. So we split the task into three parts:

- Estimating the keypoints of joints from the image
- Estimating 3D depth features and rotations
- Supervise the 3D rotation features using previous keypoints estimation

For estimating keypoints, we have a separate branch which extracts features using deconvolution layers and then regresses keypoints supervised with the ground truth keypoints in dataset. For depth and rotation features, we have another branch which also uses deconvolution layers. To make use of the keypoints information, we concatenate the keypoints features and rotations features. On top of that, we use element-wise multiplication on the keypoint features and rotation features. The idea behind this is that the keypoints features contain location information of the joints, and the rotation features should be estimated from those joints. Therefore an element-wise multiplication makes sure that the keypoints and rotations are estimating from the same image locations. We then concatenate both the products and the original features together and perform filtering. After concatenation, the features are fed into convolution blocks. Residual cuts are added between the blocks to prevent a vanishing gradient. In the end the output features is fed into a HMR-style Regressor, where we get the final predicted poses. This regressor also receives the camera and shape prediction of the simpler branch as well as recurrent loop with shared weights.

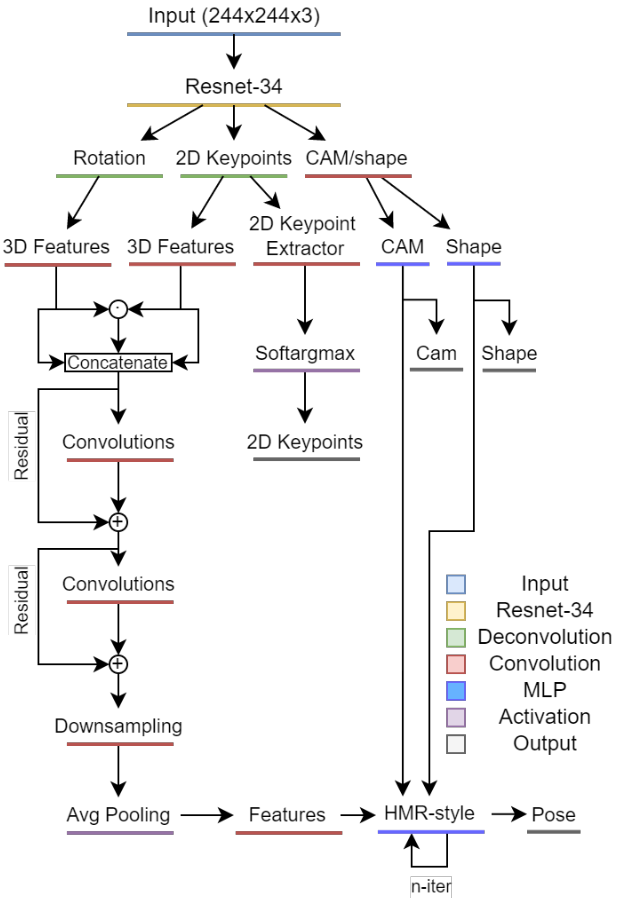

Figure 1: Architecture diagram of the model

3 EVALUATION

We experimented with variations of the PARE[1] and Hybrik[2] models, however reformating the loss and data mapping to fit the provided data and outputs proved to be exhausting. The training of our model is quite straight forward, with the main bottleneck being CUDA memory for the larger memory. We set the image size to 224 as this is the minimum requirement for the ResNet backbone. For the 2080 Ti on the Euler cluster we set the batch size to 32, while testing on a local 3090 we used a batch of 52. For the optimizer we used Adam with a learning rate 0.001. We trained the model on Euler cluster with 4 cores and GPU. The whole training takes 120 epoch, around 12 hours to converge. In the end we achieve a public score of 60 on the leaderboard.

4 DISCUSSION

Our very first approach was the Bayesian Approach for the regression. The idea behind it was that to add uncertainty to the model and make it more robust. The problem is that the training becomes much harder and takes longer because we need to estimate the gradient using Monte-Carlo Method. There was also the downside of instabilities due to sampling. Then we tried with a graph neural network for human pose, because human pose tree can be viewed as an undirected graph. This approach failed because the limbs of human bodies are extremely flexible. The pose of a body part does not contain enough information regrading its limbs and it leads to moderate poses. For data preprocessing, we tried to use augmentations such that random noises, flipping and rotations, however this reduced model convergence. We do believe that doing data augmentation can improve the performance if the hyperparameters are properly tuned, but the time was limited and this approach was abandoned.

5 CONCLUSION

The model we propose in this report is easy to understand and only uses convolution filters, which makes it intuitive and easy to train. Both training and predictions are fast and the performance in the end is also comparable with other more complicated approaches.

REFERENCES

[1] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. 2021. PARE: Part Attention Regressor for 3D Human Body Estimation. In *Proceedings International Conference on Computer Vision (ICCV)*. IEEE, 11127–11137.

[2] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3383–3393.