INTERNSHIP PROJECT REPORT

Name: Vedang Patel
External Mentor: Prof. Junsong Yuan (SUNY at Buffalo, NY, USA)
Academic Mentor: Dr. Shitala Prasad (IIT Goa)
Date: 30th Nov 2023

---

1. Introduction

Monocular 3D human pose estimation, which strives to recover 3D locations of body joints from an RGB image, has garnered significant attention and achieved impressive progress in recent years. As a longstanding task in computer vision, it continues to be challenging due to the complex articulated structure of the human body, substantial variations in poses and orientations, severe (self-)occlusions, and the inherent 2D-to-3D scale and depth ambiguities.

In this project, we expand the scope of the Neural Voting Field (NVF) framework, initially developed for camera-space 3D hand pose estimation, to the challenging field of root-relative 3D human pose estimation from single RGB images. Our adaptation applies NVF's groundbreaking 3D dense regression approach, which revolutionizes the conventional multi-stage methods. We harness NVF's dense 3D point-wise voting within the camera frustum and direct dense modeling in the 3D domain, inspired by Pixel-aligned Implicit Functions, to intricately model the human body's global geometry and local evidence. This approach is pivotal in resolving common 2D-to-3D ambiguities in human pose estimation.

In our adapted model, for each 3D query point relative to the body root in the camera frustum, NVF, empowered by a Multi-Layer Perceptron, regresses: (i) its signed distance to the nearest body surface; (ii) a set of 4D offset vectors, comprising 1D voting weight and a 3D directional vector to each body joint. A weighted average, derived from the 4D offset vectors of near-surface points, is utilized to calculate the coordinates of body joints.

Our experiments, conducted on benchmark datasets, demonstrate that the adapted NVF not only achieves state-of-the-art results in root-relative 3D human pose estimation but also retains the original framework's efficiency and accuracy. This adaptation of NVF opens new possibilities in human motion analysis, augmented reality, and interactive applications, marking a significant contribution to the advancement of human pose estimation technology.

2. Problem Statement

The primary challenge addressed in this project is the accurate and efficient estimation of root-relative 3D human poses from single RGB images. Traditional methods in human pose estimation often rely on complex, two-stage processes, which first involve either holistic or pixel-level dense regression to approximate the 3D pose relative to the camera. These are then followed by intricate secondary operations for the recovery of the global root position or scale. Such approaches are not only computationally intensive but also prone to inaccuracies, particularly in capturing the full complexity and variability inherent in human postures.

Moreover, existing techniques frequently struggle with 2D-to-3D ambiguities, failing to effectively model the dense local evidence and global geometry of the human body in three

dimensions. This limitation leads to less accurate reconstructions and an inability to faithfully represent the nuanced dynamics of human movement.

3. Proposed Solution

3.1 Key Design Elements:

3.1.1 Exploiting Dense Local Evidence:

Our method leverages dense regression-based techniques, superior for handling articulated 3D structures by maintaining spatial structure and exploiting local evidence.

3.1.2 Understanding 3D Global Geometry:

Incorporating insights from previous works, we emphasize understanding the 3D structure to resolve depth ambiguities, crucial for accurate pose estimation.

3.2 Integration with Pixel-aligned Implicit Function (PIFu) [1]:

Our methodology builds on the principles of PIFu, renowned for its efficacy in 3D human geometry reconstruction from RGB imagery. PIFu, by learning an implicit function in a 3D domain with pixel-aligned features, generates detailed 3D human geometry including largely occluded regions from a segmented RGB image. We harnessed this technique for its unparalleled ability to model detailed local features, such as clothing textures, and to provide a comprehensive representation of global geometry, crucial for accurately depicting occluded body regions.

3.3 Adaptation of NVF-In-Depth Process:

Our adaptation of NVF involves a meticulous process of analyzing 3D query points sampled within the camera's viewing frustum. For each of these points, aligned with corresponding image features, our model conducts dual regressions:

3.3.1 Signed Distance Function (SDF)[2]:

A cornerstone of our approach, the SDF computes the precise signed distance from each 3D point to the closest body surface. This calculation is vital for accurately rendering the human body's shape and depth in three dimensions, providing a clear distinction between the body's surface and surrounding space.

3.3.2 Dense Offset-based Pose Re-Parameterization:

This innovative step redefines pose estimation. We first extract a C-channel feature map F from the input RGB image by an hourglass network as used in [1, 3]. Then, for a 3D query point p sampled in camera space, we define a continuous implicit function NVF realized by an MLP, which generates a set of 4D offset vectors for each query point. These vectors consist of a singular voting weight and a tri-dimensional directional vector directed towards each body joint. This re-parameterization technique is pivotal in pinpointing the exact location of each joint, thereby enhancing the overall accuracy of the pose estimation.

3.3.3 Vote-Casting Mechanism and Joint Coordinate Calculation:

Following the regression analysis, a sophisticated vote-casting mechanism is employed. This involves selecting the 4D offset vectors from points proximate to the body surface(the set of K-Nearest Neighbors (KNN) to each joint), as indicated by their SDF values. These vectors are then aggregated using a weighted average method to calculate the precise coordinates of the body joints. This final step is crucial as it amalgamates the information gathered from multiple near-surface points to

form an accurate and coherent representation of the human pose.

4. Implementation Detail

4.1 Dataset

We use Human3.6M [4]. Human3.6M contains 3.6M video frames with 3D joint coordinate annotations. Because of the license problem, previously used groundtruth SMPL parameters of the Human3.6M are inaccessible. Alternatively, we used SMPLify-X [5] to obtain groundtruth SMPL parameters. Although the obtained SMPL parameters are not perfectly aligned to the groundtruth 3D joint coordinates, the error of the SMPLify-X is much less than those of current state-of-the-art 3D human pose estimation methods. Therefore, we think using SMPL parameters from SMPLify-X as groundtruth is reasonable. Note that for a fair comparison, all the experimental results of previous works are reported by training and testing them on our SMPL parameters from SMPLify-X.

4.2 Evaluation Metrics

MPJPE and PAMPJPE are used for the evaluation [6], which is Euclidean distance (mm) between predicted and groundtruth 3D joint coordinates after root joint alignment and further rigid alignment, respectively.

4.3 Loss function

4.3.1 L1 Loss:

L1 loss, also known as least absolute deviations (LAD). In the context of signed distance optimization, we have used L1 loss to minimize the error in estimating the signed distance from a 3D query point to the nearest surface in the model.

4.3.2 Huber Loss[7]:

Huber loss, also known as smooth mean absolute error, is a combination of L1 and L2 loss. It behaves like L2 loss for small errors and like L1 loss for large errors. This is achieved through a delta value that decides the threshold at which the loss shifts from L2 to L1. In estimating offset 3D vectors, which are crucial for determining the directional relationship between 3D points and specific body joints, we have used Huber loss.

5. Project Timeline

5.1 June - July:

In the initial phase of my internship, I studied in depth a few concepts relevant to the project, delving into the intricacies of NVF hand pose estimation and Neural Correspondence Field (NCF) [3]. My study also included understanding 3D implicit representation [1,2,8,9,10]. Additionally, I scrutinized various dense regression techniques[11,12,13]. As well as [14] and [15] approaches to mesh generation and pose estimation. Then I dive into a solid foundation for basic code structures, architecture, and functionalities.

5.2 August - September:

During these months, my focus shifted to developing a specialized preprocessing pipeline for root-relative human pose estimation using NVF, leveraging the Human3.6M dataset. This stage was pivotal in enhancing my understanding and visualization capabilities in 3D pose estimation. A critical part of this phase was generating pseudo-ground truth using the SMPLify-X model, which was instrumental in preparing and preprocessing our data. I have also modified the loss function to handle inappropriate data.

5.3 October:

I focused on training the model. Despite observing a consistent decrease in loss, indicating positive model learning and convergence, we faced a predicament. The model was not being trained effectively, as evidenced by high mean joint errors and its apparent lack of learning from the training process.

5.4 November - Present:

These last months of my internship involved researching and addressing the identified training issues. I have gotten the solution. I made further improvements to the algorithm to get more accurate 3D human pose results. I have achieved state-of-art results and the NVF model outperforms other methods. Currently, I am in the process of training on other datasets 3DPW [16] and MSCOCO [17], and publication, ensuring that every aspect of the project is meticulously documented and prepared for a wider academic and professional audience.

6. Learning Outcomes

Throughout the project, I acquired valuable insights and skills:

3D Human Pose Estimation: This deep dive into 3D human pose estimation provided me with a solid foundation in the field. I deepened my understanding of 3D implicit representation and dense regression techniques through an extensive literature review.

Insights From Relevant Work: I explored past methods like holistic and 2D dense regression, gaining insights into the evolution of 3D human pose estimation techniques and understanding the problem more precisely.

Python, Cuda, and GitHub Experience: My coding skills in Python and Cuda grew as I worked on parallel training on GPUs. Using GitHub for version control further enhanced my software development expertise.

3D Geometry and Image Processing Insights: I developed a deeper understanding of concepts like intrinsic-extrinsic parameters and coordinate calculations in different spaces, which was crucial for grasping 3D geometry.

Research Experience: This project provided me with invaluable research experience, enhancing my ability to tackle complex problems and contribute to the field of 3D human pose estimation.

7. Project Deliverables

Our experiments with the NVF framework showcased its superior performance over current state-of-the-art algorithms in root-relative 3D human pose estimation on the Human3.6M dataset. While the goal of publishing a research paper was not achieved yet, this outcome was due to the extensive scope and depth of the experiments. We are optimistic about future opportunities to present our findings at conferences, following further experimentation with additional datasets.

8. References

[1] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In ICCV, 2019.

[2] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven

Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In CVPR, 2019.

[3] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In ECCV, 2022.

[4] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. TPAMI 2014.

[5] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR 2019.

[6] Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In: ICCV 2019.

[7] Peter J Huber. Robust estimation of a location parameter. Breakthroughs in statistics, 1992.

[8] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In CVPR, 2019.

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In CVPR, 2019.

[10] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In 3DV, 2020.

[11] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In CVPR, 2018.

[12] Liuhao Ge, Zhou Ren, and Junsong Yuan. Point-to-point regression pointnet for 3d hand pose estimation. In ECCV, 2018.

[13] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In ECCV, 2018.

[14] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Imageto-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In ECCV, 2020.

[15] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In CVPR, 2021.

[16] von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using imus and a moving camera. In: ECCV 2018.

[17] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV 2014.