ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 18, Issuse.1 March 2024

HUMAN 3D POSE ESTIMATION USING DEEP LEARNING

Talari Sreekanth Kumar[1],Mamidi Abhinav Reddy[2] ,Saikam Pavan Teja[3],A Naveen Kumar[4]

[1,2,3] B.Tech Student, Department of CSE (Data Science), Malla Reddy College of Engineering and Technology, Hyderabad, India.

[4]Associate Professor, Department of CSE (Data Science), Malla Reddy College of Engineering and Technology, Hyderabad, India.

***Abstract*—** Human 3D pose estimation is a fundamental problem in computer vision and has garnered significant attention in recent years due to its broad range of applications. This paper presents a comprehensive study of state-of-the-art deep learning techniques for human 3D pose estimation from 2D images or videos. The primary objective is to provide an in-depth analysis of the advancements, challenges, and future directions in this exciting field. We begin by reviewing the historical context of human pose estimation, tracing the evolution of techniques from traditional computer vision methods to the current dominance of deep learning-based approaches. We discuss the underlying concepts of pose representation, joint detection, and the importance of data annotation, emphasizing the pivotal role of large-scale datasets in training robust deep models. Our work delves into the architectural choices for 3D pose estimation networks, comparing various convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models. We explore the trade-offs between accuracy and efficiency and highlight the influence of architectural design on the overall performance.

***Keywords*-3DPose, 2D Pose,Deep Learning, Convolutional Neural Network.**

I. INTRODUCTION

In the realm of computer vision, the quest to decipher human motion and pose from 2D images or videos is transforming the way we interact with technology. Human 3D pose estimation, the art of uncovering the intricate dimensions of body movement, has leaped forward with the power of deep learning. This project delves into the captivating world of Human 3D Pose Estimation, exploring the innovative potential of deep neural networks and their impact on fields like robotics, healthcare, and more.

This project embarks on a captivating journey into the realm of Human 3D Pose

Estimation, with a keen focus on harnessing the transformative power of deep neural networks. Our aim is not only to unravel the complexities of this technology but also to showcase its profound implications across a spectrum of domains, ranging from the precise choreography of robots to the personalized care in healthcare, and even the immersive experiences of the entertainment industry.

As we delve deeper into this exploration, we will dissect the core methodologies, delve into the architectural intricacies, and uncover the pivotal role of datasets and annotations in training these intelligent systems. Our journey will also venture into the subtleties of training strategies, optimization, and the critical evaluation metrics that underpin the accuracy and usability of 3D pose estimation models.

However, to make more accurate predictions about human behaviours, we need more than a few body key points. To that end, 3D whole-body pose estimation aims to detect face, hand and foot key points in addition to the standard human body key points of classical 3D human pose estimation.

II. LITERATURE REVIEW

Transition to 3D

- The transition from 2D to 3D was accelerated by works like "VNect" (Mehta et al., 2017), which used CNNs to predict 3D poses from 2D joint locations.

Benchmark Datasets

- Datasets like "Human3.6M" and "MPII Human Pose" provided annotated data, enabling model training and evaluation.

Multi-Stage Networks

- "Stacked Hourglass" (Newell et al., 2016) introduced multi-stage networks, significantly improving accuracy.

Transfer Learning

- Transfer learning with pre-trained models on datasets like ImageNet reduced data requirements and improved convergence.

Recent Trends

- Transformer-based models, such as "Vision Transformer" (ViT), have gained prominence for capturing spatial dependencies.

III. METHODOLOGY

we describe the making of the H3WB dataset. Our objective is to build a key point-based 3D whole-body dataset including key points on the body, the face and the hands, and propose a benchmark. We use the same key point layout as COCO Whole Body dataset with 133 key points.

To that end, we build on the widely used Human3.6M dataset for which we provide 3D whole-body key points. The H3WB building process is as follows: First, we use an off-the-shelf 2D whole-body detector combined with multi-view reconstruction to obtain an initial set of incomplete 3D whole-body key points. Next, we implement a completion network to fill in the key points missed by the multi-view geometric approach. Then, we develop a refinement method for the hands and the face to obtain more accurate key points. Finally, we perform quality assessment to select 25k 3D whole-body poses with high confidence and the 100k associated images from 4-view.

Dataset	Size	Keypoints	Body	Hand	Face
Human3.6M[36]	3.6M	17	17		
3DPW[71]	51k	24	24		
LSP[41]	10k	14	14		
3DHP[71]	> 1.3M	17	17		
Panoptic[42]	1.5M	15	15		
MTC[77]	834K	20	20		
InterHand2.6M[57]	2.6M	21		21	
FreiHAND[85]	37k	21		21	
RHD[84]	44K	21		21	
MTC[77]	111K	21		21	
TotalCapture[43]	1.9M	127	21	16+16	74
ExPose[18]	33K	144	25	15+15	89
H3WB	100k	133	23	21+21	68

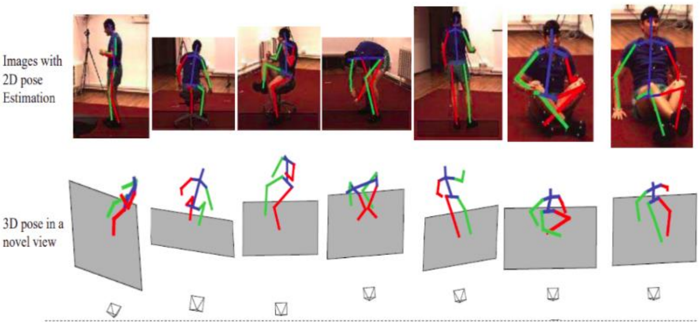
Table1.Over view of Human3.6M Dataset.

1. Convolutional Neural Network:

- Create a neural network model by extracting the features from the 2D image or video.
- Design the model architecture with multiple dense layers, incorporating appropriate activation functions such as ReLU and sigmoid.
- Compile the model with Adam optimizer, binary cross-entropy loss function, and accuracy as the metric.

2. Training the initial model:

- Train the model using the training set.
- Specify the number of training epochs and batch size to iterate over the dataset and update the model's parameters.



The occluded or undetected key points (cyan key points) are reprojections after 3D multi-view reconstruction. Notice that these reprojections do not always align with the images, like the right hand in the last view, which is probably due to OpenPifPaf not being perfectly accurate.

we use an off-the-shelf 2D whole-body detector combined with multi-view reconstruction to obtain an initial set of incomplete 3D whole-body key points. Next, we implement a completion network to fill in the key points missed by the multi-view geometric approach. Then, we develop a refinement method for the hands and the face to obtain more accurate key points. Finally, we perform quality assessment to select 25k 3D whole-body poses with high confidence and the 100k associated images from 4-view.

Initial 3D whole-body dataset with OpenPifPaf We run the 2D whole-body detector from OpenPifPaf [47] on all the 4 views from the training set of Human3.6M. Since the cameras of Human3.6M are well calibrated, we can reconstruct key points in 3D using standard multi-view geometry.

The OpenPifPaf 2D whole-body detector can miss key points due to self-occlusions (hands, feet) or unfavourable camera viewpoints (facial landmarks). However, the four-view setup allows us to recover missing key points and obtain a complete 3D whole-body pose, provided each key point appears in at least two non-opposing views.

An example of this process is shown in Figure 2. Using this method, we obtained 11,426 fully complete 3D whole-body poses with all 133 key points and 26,333 incomplete 3D whole body poses where all key points appear in at least one view, resulting in a total of 37,759 3D whole-body poses with each key point appearing in at least one view.

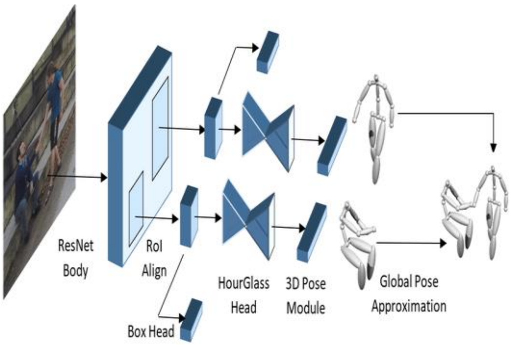


Fig1.OpenPifPaf detects most of the non-occluded key points inside the image (orange key points).

IV. IMPLEMENTATION

1. LOAD THE HUMAN3.6M DATASET USING PANDAS AND PERFORM DATA EXPLORATION.

2. PREPROCESS THE DATASET

3. SPLIT THE DATA SET INTO A TRAINING SET AND TEST SET

4. BUILD A NEURAL NETWORK MODEL USING THE
5. CNN AND OpenPifPaf.

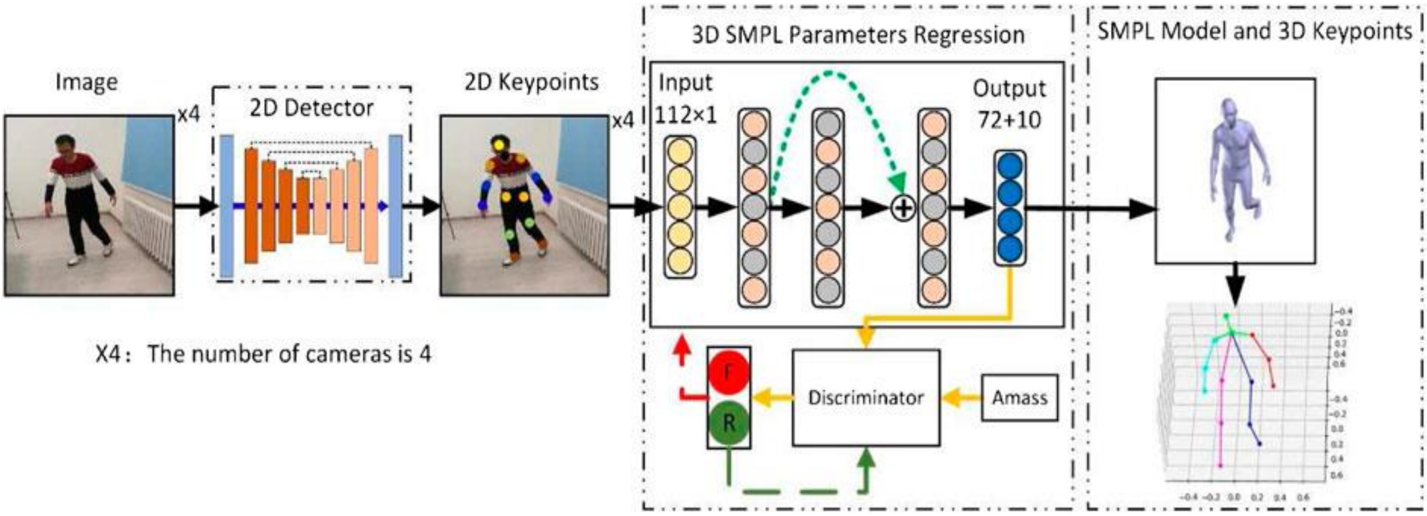


Image
x4
2D Detector
2D Keypoints
x4
3D SMPL Parameters Regression
Input
112×1
Output
72+10
F
R
Discriminator
Amass
SMPL Model and 3D Keypoints
X4：The number of cameras is 4

Fig2:Architecture

We use the H3WB dataset to propose a benchmark and the associated leader board. We split the dataset into training and test sets. The training set contains all samples from S1, S5, S6 and S7, including 80k {image,2D,3D} triplets. The test set contains all samples from S8, including 20k triplets. The test set labels are retained to prevent involuntary overfitting on the test set. Evaluation is accessible only by submitting results to the maintainers. We do not provide a validation set. We encourage researchers to report 5-fold cross-validation average and standard deviation (see supplementary).

The corresponding benchmark has 3 different tasks:

- 3D whole-body lifting from complete 2D whole-body skeletons, or 2D→3D for short.
- 3D whole-body lifting from incomplete 2D whole body skeletons, or I2D→3D for short.
- 3D whole-body skeleton prediction from image, or RGB→3D for short.

For each task, we report the following MPJPE (Mean Per Joint Position Error) metrics:

i.MPJPE for the whole-body, the body (key point 1-23), the face (key point 24-91) and the hands (key point 92133) when whole-body is cantered on the root joint, i.e. aligned with the pelvis, which in our case is the middle of two hip joints,

ii.MPJPE for the face when it is cantered on the nose, i.e. aligned with key point 1, iii.MPJPE for the hands when hands are cantered on the wrist, i.e. left hand aligned with key point 92 and right hand aligned with key point 113.

We propose a second task where we want to obtain 3D complete whole-body poses from 2D incomplete pose. This task aims to simulate the more realistic

case when there are occlusions and the 2D whole-body detector outputs an incomplete skeleton. We do not provide masks for the training skeletons to allow for online data-augmentation.

Instead,we propose a masking strategy as follows:

1. With 40% probability, each keypoint has a 25% chance of being masked,
2. with 20% probability, the face is entirely masked,
3. with 20% probability, the left hand is entirely masked,
4. with 20% probability, the right hand is entirely masked.

• ethod	MAll	Body	Face / aligned†	Hand / aligned‡
H3WB				
SMPL-X[60]	188.9	166.0	208.3 / 23.7	170.2 / 44.4
CanonPose[72]*	186.7	193.7	188.4 / 24.6	180.2 / 48.9
SimpleBaseline [53]*	125.4	125.7	115.9 / 24.6	140.7 / 42.5
CanonPose[72] *w* 3D sv.*	117.7	117.5	112.0 / 17.9	126.9 / 38.3
Large SimpleBaseline[53]*	112.3	112.6	110.6 / 14.6	114.8 / 31.7
Jointformer[52]	88.3	84.9	66.5 / 17.8	125.3 / 43.7
H3WB+T3WB				
CanonPose[72]*	164.7	161.1	174.5 / 21.5	150.8 / 43.6
SimpleBaseline [53]*	115.3	114.8	109.4 / 15.8	125.1 / 33.5
Jointformer[52]	81.5	78.0	60.4 / 16.2	117.6 / 38.8

Table2.Comparing different methods for I2D→3D on H3WB test set.Results are shown for the MPJPE metric in mm.

Multiple Persons. Compared with single human pose estimation, estimating 3D poses of multiple persons is more challenging. When estimating multi-person from a monocular image, the additional challenge is the occlusion caused by nearby individuals. When estimating 3D poses of multiple persons from multiple views, the main challenges include the larger state space, occlusions and cross-view ambiguities, as shown in Fig. 2. Besides, most existing methods are based on two-stage frameworks which suffer from problems in efficiency, while single-stage methods (Nie et al., 2019) have been proposed to solve this problem, they are far from mature.

Triangulation is another fundamental method for reconstruction in computer vision. EpipolarPose (Kocabas et al., 2019) uses the epipolar geometry method to recover the 3D pose from the 2D poses and uses it as a supervision signal to train the 3D pose estimation model, as shown in Iskakov et al. (2019) first propose a baseline method that feeds the 2D joint confidences and 2D positions of all views produced by the 2D pose detector to the algebraic triangulation module to obtain the 3D pose. The drawback of this method is that images from different cameras are processed independently. Therefore, a more powerful triangulation procedure is proposed by them. During processing, the features maps are not projected into 3D volumes and the volumes from multiple views are aggregated and processed by a 3D CNN to output 3D heatmaps.

SMPL-Based Model: For the shape model, recent works use the skinned multi-person linear (SMPL) model (Loper et al., 2015), as shown in Fig. 6, to estimate 3D human body joints (Bogo et al., 2016). The human skin is represented as a triangulated mesh with 6890 vertices, which is parameterized by shape and pose parameters. The shape parameters are used to model the body proportions, height and weight, while the pose parameters are used to model the determined deformation of the body. The 3D pose positions can be estimated by learning the shape and body parameters.

We use statistics from the training set to adjust the test predictions. We calculate a scaling factor using the ratio of 3D to 2D bounding boxes. The formula is:

Xfinal = Xunit ×

where Xunit is the normalized prediction, σ3d is the average size of the 3D training boxes, σ2d is the size of the current 2D box, and σ2d is the average size of the 2D training boxes.

Since Simplify-X has 144 key points with a different layout, we use interpolation to transform between the Whole-Body skeleton and SMPL-X and run SMPL-X's optimization for 2,000 iterations (4 minutes/sample).

We present the results in SMPLify-X performs the worst, showing that parametric models struggle more than discriminative approaches. Simple Baseline is a solid method, and Large Simple Baseline improves its performance further. Canon Pose can be improved with additional 3D supervision, but still performs worse than Large Simple Baseline. Canon Pose also predicts the camera view, and the uncertainty in this prediction can lead to more error. Joint former achieves the best results among all methods, but still has room for improvement. All methods perform worse on our benchmark than on Human3.6M because of pelvis cantering, which creates higher numerical error on extremities like hands and face, the parts that contain most of the whole-body key points.

H3WB annotations

To download the H3WB dataset annotations click here. The zip file contains following:

- 2Dto3D train.json has the training annotations for 2D→3D and I2D→3D tasks. Since this file is too big, we split it into 4-parts to ease the training and data loading pipeline. We provide the splitted files as well.
- RGBto3D train.json has the training annotations for RGB→3D task.
- 2Dto3D test 2d. json and I2Dto3D test 2d. json include test instances for 2D→3D and I2D→3D tasks, respectively.
- RGBto3D test img.json includes test samples for RGB→3D task.

3DPW (3D Poses in the Wild, von Marcard et al. (2018)) is the first dataset in the wild with accurate 3D poses for evaluation. It is created by utilizing information from IMUs and a hand-held phone camera. A 3D pose estimation method named video inertial poser (VIP) is used to integrate the images and IMU readings of all frames in video sequences. The VIP has been validated on the Total Capture dataset, which has an accuracy of 26 mm and is accurate enough to create the dataset for image-based 3D pose estimation. For tracking single subjects, 17 IMUs would be used, while 9–10 IMUs would be used to simultaneously track up to 2 subjects. Then, the video and IMUs data are synchronized by a clapping motion as in Pons-Moll et al. (2011). In total, the dataset contains up to 18 clothing styles and actions such as walking in cities, going up-stairs, having coffee, or taking the bus. Compared with Total Capture, there are more subjects in a scene.

V.RESULTS

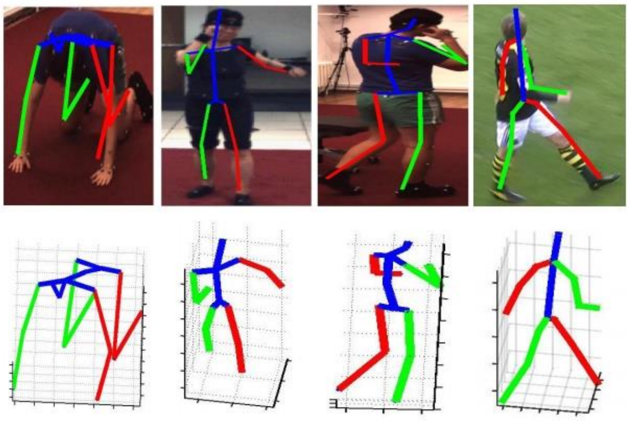


Fig3.3d Pose from 2d.

VI.CONCLUSION

In this paper, we introduce the H3WB dataset, which extends the Human3.6M dataset with 2D and 3D keypoint annotations for body, face, and hands, containing 100k images with 133 keypoints with an average accuracy of 17mm. We propose three tasks based on this dataset: 3D wholebody lifting from complete 2D keypoints, 3D whole-body lifting from incomplete 2D keypoints, and 3D whole-body prediction from monocular images. We evaluate several baselines on these tasks and demonstrate promising accuracy, but with room for improvement. Lifting from incomplete 2D skeletons and direct estimation from monocular images remain challenging, and we hope that our dataset and benchmark will spur future research in these areas.

Human 3D Pose Estimation is a computer vision and deep learning task that involves determining the three-dimensional positions of key anatomical points or joints on the human body from two-dimensional images or video frames. It is a fundamental problem with numerous applications in fields such as robotics, healthcare, sports analytics, entertainment, and more.

The primary goal of Human 3D Pose Estimation is to reconstruct the 3D positions of human body joints or landmarks in a coordinate system that corresponds to the physical world. These joints typically include key points like the head, shoulders, elbows, wrists, hips, knees, and ankles. The result is a representation of the human body's posture and movement in a 3D space.

2D Pose Estimation: In the initial stage, a 2D Pose Estimation model is employed to detect and localize body joints in a 2D image or video frame. This step provides the 2D positions (x, y) of the joints.

Depth Estimation: To convert 2D joint positions into 3D coordinates, depth information is needed. Depth estimation techniques, such as stereo vision or depth sensors (e.g., RGB-D cameras), are used to infer the depth (z-axis) of each joint.

3D Pose Reconstruction: With the 2D positions (x, y) and depth information (z), the 3D pose of the human body is reconstructed. This involves transforming the 2D coordinates into a 3D space using geometric principles.

VII.REFERENCES

- ➢ Dennis Bautembach, Iason Oikonomidis, and Antonis Argyros. Filling the joints: Completion and recovery of incomplete 3d human poses. *Technologies*, 2018.
- ➢ Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999.
- ➢ Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. *ECCV*, 2016.

- Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019.
- Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, 2018.
- Mercedes Garcia-Salguero, Javier Gonzalez-Jimenez, and Francisco-Angel Moreno. Human 3d pose estimation with a tilting camera for social mobile robot interaction. *Sensors*, 2019.
- Erik Gartner, Aleksis Pirinen, and Cristian Sminchisescu.¨ Deep reinforcement learning for active human pose estimation. *AAAI*, 2020.
- Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, pages 10833–10842, 2019.
- Yiwen Gu, Shreya Pandit, Elham Saraee, Timothy Nordahl, Terry Ellis, and Margrit Betke. Home-based physical therapy with an interactive computer vision system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.
- Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, Jose MF Moura, and Manuela Veloso.´ Teaching robots to predict human motion. In *IROS*, 2018.
- Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. *CVPR*, 2019.
- Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Handsformer: Keypoint transformer for monocular 3d pose estimation ofhands and object in interaction. *CVPR*, 2022.