DEGREE PROJECT IN TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2020

3D Human Pose Estimation

KTH Thesis Report
Data, Methodology and Results
Draft

Sri Datta Budaraju

KTH ROYAL INSTITUTE OF TECHNOLOGY
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Authors

Sri Datta Budaraju <budaraju@kth.se>
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden
Stuttgart, Germany

Examiner

Danica Kragic Jensfelt
Stockholm, Sweden
KTH Royal Institute of Technology

Supervisor

Hedvig Kjellström
Stockholm, Sweden
KTH Royal Institute of Technology

Supervisor - Host

Arij Bouaziz

Stuttgart, Germany

Mercedes-Benz AG, Research and Development

Acronyms

AR/VR Augmented Reality/Virtual Reality

HPE Human Pose Estimation

RGB Red Green Blue

VAE Variational Auto-Enocder

MVAE Multimodal Variational Auto-Enocder

MMVAE Mixture-of-Experts Multimodal Variational Auto-Enocder

SOTA State-of-The-Art

POV Point of View

NRSfM Non-Rigid Structure from Motion

MoCap Motion Capture

KLD Kullback–Leibler Divergence

L1 Least Absolute Deviations

MSE Mean Squared Error

MPJPE Mean Per Joint Position Estimate

UMAP Uniform Manifold Approximation and Projection

Contents

Chapter 1

Introduction

With rapid advancements in deep learning facilitated by the developments in computational hardware, there has been tremendous growth in computer vision research and its applications [4]. One of the major tasks of computer vision that is required for real-world applications is to perceive and understand dynamic objects and more importantly humans.

Human Pose Estimation, also referred to as HPE is a fundamental problem in computer vision that also forms a basis for human action and gesture recognition as well as human motion prediction. Human Pose Estimation (HPE) is defined as the localization of human joints (also known as keypoints, including head, eyes, ears, nose etc) mainly in images and videos in either a 2D or 3D coordinate space. The widely available and used data like images and videos are 2-dimensional data and lack spatial information which is crucial for most of the applications like autonomous driving, Augmented Reality/Virtual Reality (AR/VR), social robots etc. Hence this thesis focus is on 3D Human Pose Estimation.

1.1 Background

There has been a lot of research done in 3D human pose estimation and more advancements have been made in the past few years leveraging the power of deep learning. The current state of the art methods explores various ways to solve the task using Red Green Blue (RGB)/Depth image channels, 2D poses, 3D poses, multi-view and sequential images. The main State-of-The-Art (SOTA) approaches either directly estimate 3D pose from an RGB/D image in a bottom-up manner, or start with an intermediate 2D pose or, 2D joint heatmaps to finally recover the 3D pose by a *lifting* network. Typically, these approaches directly estimate the 3D coordinates of the keypoints or estimate the shape and camera parameters to reconstruct the 3D pose.

The former approaches are usually trained in a cascaded manner i.e by having an intermediate state that learns 2D pose in some way. Most of these methods have complex architectures that are hard to train or use multi-view images making it impractical to scale the training to the wild, where such data is very hard to obtain. Since 2D poses are naturally obtained by projecting 3D poses to a plane, the latter approach of lifting 2D-to-3D is an *ill-posed inverse* problem due to its inherent ambiguity.

Non-supervised (Weak/Semi/Self/Un-supervised) learning regimes have also gained traction in 3D HPE recently and many of the deep learning techniques that have already improved the results in other computer vision tasks (and even in supervised HPE), are yet to be explored.

1.2 Problem

How can we learn a strong visual representation of the data to tackle the 3D-pose estimation? Could data as its own supervisory goal (self-supervision) resolve the ambiguities of the pose estimation?

1.3 Goal

The main aim of the thesis is to investigate SOTA unsupervised learning approaches to estimate 3D human pose from images. And to also investigate 2D-to-3D lifting methods to tackle the challenges of 3D pose estimation in the wild.

Improvements in the aspects of ease of training procedure i.e requiring less data or less labor-intense labeling, inference speed, and most importantly accuracy is important and will directly impact its super tasks such as, action and gesture recognition, motion prediction and intention/behaviour prediction.

1.4 Benefits, Ethics and Sustainability

Human Pose Estimation plays a very important role to enable autonomous vehicles and robots to safely interact with humans. It plays a key role in developing higher dimensional communication platforms with AR/VR. It is crucial for surveillance systems to ensure public safety. However such important technologies are only as good as the intentions of its users. Mass surveillance of citizens by their governments is a matter of debate.

1.5 Methodology

The problem of 3D HPE has 3 aspects to be addressed and explored.

The neural network: The architecture and the kind of neural network to be used. 3D poses can be predicted using regular linear neural network, or using various forms of autoencoder architectures. These models can use linear, convolutional or graph networks to learn the features. This thesis focus on exploring all the above-mentioned kinds of networks to solve the 3D HPE within the context of probabilistic inference models, typically VAEs, as a deterministic approach for an inherently ill-posed problem is not ideal.

The learning task: The model could either learn to directly predict the 3D coordinates of the keypoints, or learn structural parameters that could model a 3D pose. The thesis only explores the former task.

The learning technique (or the cost): . The model can be either trained by directly comparing the predicted 3D pose and the ground truth thus requiring 3D annotations or, by projecting the prediction back to 2D to compare with the input (requires only 2D annotations that could be acquire from SOTA in 2D HPE). Adversarial training and self-supervision techniques have also given promising results in the last couple of years. The thesis's prime focus is on the first aspect of investigating the merit in using VAE for 3D HPE hence direct comparison of 3D pose and ground truth would be used and could be further extended to other techniques with moderate modifications.

1.6 Stakeholders

Daimler's 'Environment Perception for Autonomous Driving' R&D team in Stuttgart conducts cutting-edge research in the field of Computer vision and Deep Learning to improve the State-Of-The-Art and to make Autonomous Driving a reality. This thesis is part of the team's on-going research in the area of Human Pose Estimation which would help autonomous cars better perceive, understand and interact with humans. Daimler/Mercedes-Benz autonomous cars try to understand humans both, inside and outside the car and Human Pose Estimation is a critical element to accomplish the task.

The question is also of interest to the research area of Human State/Action Recognition in specific and also to areas of computer graphics to model humans in 3D space. Hence it is beneficial to various areas that try to understand and interact with humans. The scientific communities in the areas of Autonomous Driving, AR/VR, Motion Capture, Computer Graphics, and Human-Robot interaction could be interested in the contribution of this thesis.

1.7 Delimitations

This thesis focuses only on 3D pose estimation and not the intermediate 2D pose. Data collection is not part of the thesis study but uses only publicly available, widely used and benchmarked datasets.

1.8 Outline

The current version of the draft consists of two chapters alone. Chapter 2, the Theoretical Background further contains related works and would later include theoretical concepts that would be touched in the methodology.

Chapter 2

Theoretical Background

In this chapter the details of various related works and the state-of-the-art in 3D Human Pose Estimation are presented along with some works in the sibling task of hand pose estimation.

2.1 Related Work

2.1.1 Pose from images

There are numerous works that try to estimate 3D human poses from 2D RGB images or 2D joint confidence heatmaps [1, 3, 16, 22]. Most of these methods follow a cascading approach, where an explicit intermediate representation of 2D pose or 2D heatmaps is used.

For example, [16] proposes a general framework with 3 networks. Human detection Network, RootNet, PoseNet. Where, the human detection network predicts the region the human is in an image. The RootNet localizes the human's root in the global 3D world. And, the PoseNet predicts the 3D pose of a single person with respect to the root. Where, the root is a fixed reference point of the human body say, pelvis.

The advantage of such top-down frameworks is to divide the task of RGB to 3D into smaller, well-studied sub-tasks. This makes scaling single-person pose estimation algorithms for multi-person pose estimation easy, as the majority of the data available mostly consists of a single person per frame. In addition to it, this approach provides the opportunity to improve certain modules without affecting or having to re-train the other modules of the system.

2.1.2 Pose Lifting

In contrast to the estimating pose from an image, Pose Lifting works such as [1, 2, 13, 17, 21], focus on estimating 3D poses from 2D poses alone. Assuming 2D poses from the SOTA methods in 2D HPE. These methods include simple linear models as first described in [15] with a series of fully connected linear layers, and sometimes batch normalization, dropout and, residual connections to regress 3D pose effectively.

Non-Rigid Structure from Motion (NRSfM) is another promising lifting method that also leverages images along with 2D annotations. NRSfM deals with the problem of reconstructing 3D shape (pose/point cloud) and cameras of each projection from a sequence of images with corresponding 2D orthogonal projections (2D keypoints). This approach has been widely used in facial keypoint detection and [12] introduces deep learning variant for the same. Instead of predicting the 3D coordinates of each keypoint/joint of the 3D pose, [12, 17, 22, 23] predicts the 3D shape and camera pose from 2D pose using this method.

The Pose Lifting approach facilitates to leverage the already well established 2D HPE models that are trained on enormous and diverse labelled data. Thus demanding lesser training data for 3D pose estimation than it would need when learning from images. Since these networks do not have large convolution layers they are less computationally inexpensive for both training and inference on edge computing units. Moreover, the 2D and 3D pose data usually can be entirely loaded on to GPU further accelerating the training procedure. Thus addressing the critical problem that hinders scalability of 3D HPE models and also helps to develop better modular systems by combining the best of Pose Lifting networks with the best of 2D HPE. However, due to the inherent ambiguity in lifting pose to 3D and as the images are not captured with orthogonal cameras, reprojection of 3D pose will vary from the ground truth, it is challenging to match the performance of models trained on 3D ground truth.

2.1.3 Non-Supervised Learning

The standard way to train 3D/2D HPE is by minimizing the distance between the predicted 3D/2D pose and its corresponding ground truth. The area of 2D HPE is well established and matured with reliable systems deployed in the real world. This was made possible with the high volume of images from diverse settings and the reasonable ease of manual labelling of 2D poses. On the other hand, labelling 3D pose manually is not practical. Though single-person datasets such as Human3.6M [11], Human Eva [10] and, multi-person datasets such as CMU Panoptic [8] provide 3D pose ground truth. They are obtained using Motion Capture (MoCap) systems which are only limited to indoors or cannot be directly adapted to outdoor environments where the majority of the use cases exist fig[2.1.1]. It is also worth mentioning JTA(Joint Track Auto) dataset [5] that is made using the GTA(Grand Theft Auto) game engine which is technically scalable with its own limitations. But datasets from simulations come with the difficulty of domain adaptation to be transferable to the real world.

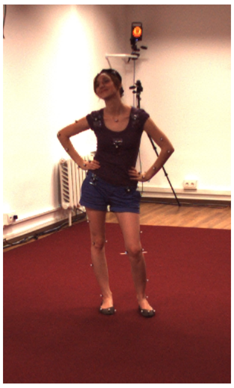

Figure 2.1.1: Image from Human3.6 Dataset [11] of subject wearing MoCap markers

To overcome this bottleneck, [13] proposes unsupervised training of a generative adversarial network by projecting the predicted 3D pose back to 2D and minimizing its distance with the input 2D pose. And further training a discriminator to distinguish the real 2D pose from the projected poses. Thus removing the need for any explicit 3D annotations besides 2D poses that are either manually labelled or obtained using 2D HPE models. RepNet [21], trains an adversarial network without 2D-3D correspondences in a weakly supervised manner. Moreover, it also does not require camera parameters to project the 3D pose but learns to predict them. Thus enabling better generalization to more diverse data with unknown cameras and poses.

To test the maximum capability of Pose Lifting networks, [2] proposes a combination of unsupervised and adversarial learning that mainly leverages the property of *plane-invariance*. It is the property that 2D projections of a 3D pose from different camera viewpoints, when lifted should produce identical and the original 3D pose. In this method, the predicted 3D pose is rotated in random angles and is reprojected to 2D in a different Point of View (POV). A discriminator is then used to evaluate if this new 2D pose is in the possible pose distribution which is learnt from 2D pose datasets alone. These steps are redone in reverse order to obtain the original 2D input. This cycle provides three intermediate representations of the single 2D input that the models learn from. Additionally, this approach exploits the temporal consistency in the datasets as well as integrates a domain adaptation network to learn from different datasets and distributions to achieve comparable results to that of the methods that require more supervision.

2.1.4 Multimodal Representation Learning

Another interesting approach is training VAEs using multiple modalities like images, poses, depth maps [7, 18–20]. Multimodal Variational Auto-Enocder (MVAE)s learn representation from different modalities in the same latent space. True multimodal learning needs to fulfill 4 criteria as follows: i) *Latent Factorization* - Implicit factorization of latent space into private, shared subspaces based on modality as illustrated in the figure[2.1.2]. ii) *Coherent Joint Generation* - Coherence in generations of different modalities from the same latent value with respect to the shared aspects of the latent. iii) *Coherent Cross Generation* - Generation of one modality conditioned on data from different modality while preserving the similarity between them. iv) *Synergy* Enhancement in generation quality of one modality as a result of learning representations of different modalities.

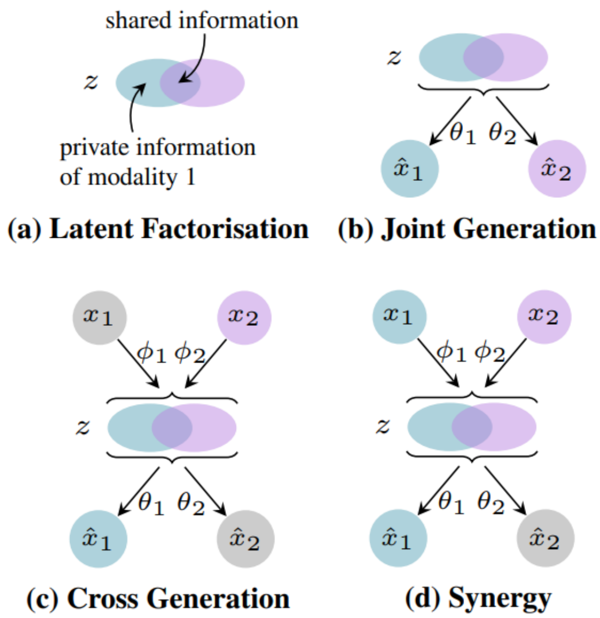

Figure 2.1.2: Criteria for Multimodal Generation [18]

Mixture-of-Experts Multimodal Variational Auto-Enocder (MMVAE) proposed by [18] fulfills all 4 of the above-mentioned criteria learning representations of image and text data, while other approaches focus on leveraging specific advantages of multimodal learning. Consider the cross-modal learning for 3D Hand Pose Estimation proposed by [19]. It involves training an encoder-decoder pair to learn image representation, and another such pair to learn 3D hand pose representations in the same latent space. This training procedure focuses on cross-generation and synergy. That is, using the shared latent space of the image and pose representations, the RGB image encoder combined with the pose decoder can generate 3D poses and vice versa while preserving the commonality between the conditioned and the generated data. With this approach, it is possible to train a VAE for 3D HPE from RGB images without explicit intermediate stages like the earlier mentioned cascading approaches. Making it more efficient and fast for both training and inference without compromising the modularity offered by cascading approaches.

Chapter 3

Data

This chapter discusses the datasets used in the thesis, as well as the processing steps to make the data more learnable. The main dataset used in the thesis is *Human3.6M*: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments [11]. Most of the related works benchmark their methods on Human3.6M and it also is freely accessible to academics on request. For further evaluation of model performance in the wild, outdoor datasets that do not have 3D ground truth such as *3DPW*: 3D Poses in the Wild [14] would be used.

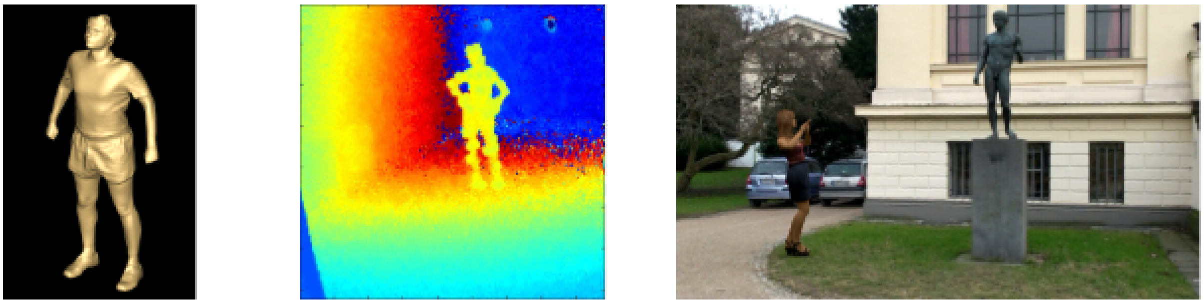

Figure 3.0.1: Full body model, depth from time of flight and mixed reality in Human3.6M dataset

3.1 Human3.6M

Human3.6M is a large scale indoor dataset with 3.6 million human poses collected with 4 cameras at different angles using a highly accurate maker-based MoCap system. The dataset constitutes 15 diverse motion and actions such as eating, sitting, walking in various everyday scenarios such as, a hand in the pocket, talking over the phone, walking a dog etc. These actions are performed by 11 professional actors wearing a variety of realistic clothing. The datasets provides synchronised 2D and 3D data including full-body scans as shown in figure[3.0.1]. It also includes mixed-reality test data created using animated human models to cover huge variations of background, clothing, illumination, occlusion and camera angles.

3.2 Processing

The methods explored by this thesis would require only images, 2D and 3D human pose from the dataset. The following are the pre-processing steps for the 2D and 3D poses.

The 3D pose in the dataset that are obtained from the marker-based MoCap are in a global reference frame. These poses using the camera parameters, are transformed into the camera coordinate frame. For the task of predicting 3D pose from either images or 2D pose, it is unrealistic to directly estimate all the joints of the pose in a global frame. So the first step of processing would be to zero the pose w.r.t the root joint say, Pelvis. As the root is always zero, we remove it so we do not have to learn the constant joint. Removing Pelvis, 16 out of the 17 joints or keypoints remain. The 3D pose is then normalized with the mean and standard deviation of the entire training and validation poses respectively.

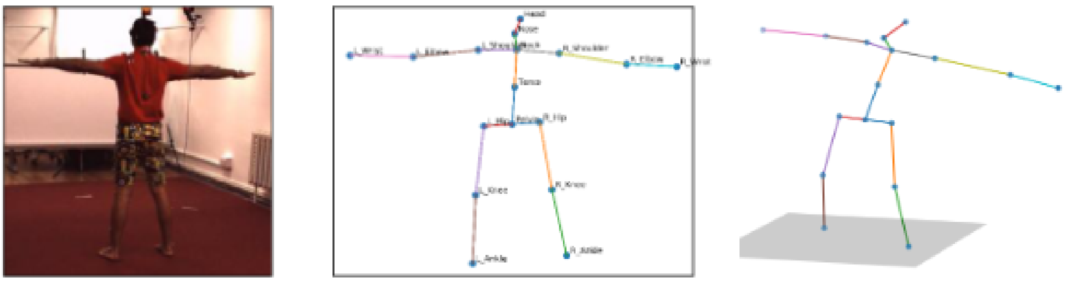

Figure 3.2.1: Human3.6M Pose Sample

The 2D pose which is obtained from the 3D pose, is also in the camera coordinate frame. The 2D pose is not zeroed as the cameras used to capture the data are not orthogonal but are perspective cameras. Zeroing the root of the 2D would eliminate the perspective information that could be important to estimate the 3D pose. The experiments to test the importance of that information is yet to be done. The root of the 2D pose is however removed to remain consistent with the 3D pose. Similar to the 3D, the 2D pose is also normalized. An image sample from the dataset with its corresponding 2D and 3D pose before normalization are illustrated in the figure[3.2.1].

The estimated poses from the networks that are trained on normalized poses are denormalized to the original scale using the same mean and standard deviation. This postprocessing step is required for getting the distance between prediction and ground truth keypoints in understandable units like millimeters. It is also necessary to denormalize for visualizing the poses, as poses normalized over all the keypoints appear heavily skewed and distorted.

Chapter 4

Method

The initial method that is being explored is using VAE to estimate 3D pose from both RGB image and 2D pose. This approach is based on the crossmodal hand pose estimation [19] but with the goal to test the performance on human poses and to investigate other ideas and techniques that the paper has not addressed. Currently we explore the image and pose modalities and investigate training VAE for image to 3D and 2D to 3D pose estimation.

4.1 CrossModal Architecture

As described in section[2.1.4], the crossmodal training involves training the encoders of each modality to learn to represent the input in the same latent space. Similarly the decoders learn to sample an embedding from this shared latent space and reconstruct an image or pose respectively. In contrast to the [19], that uses an RGB to RGB and 3D to 3D encoder-decoder pair to make enable self-supervision, we use 2D encoder instead of a 3D to evaluate cross-generation and synergy for 3D HPE from images or 2D pose. The prediction of the 3D decoder could be reprojected to 2D to eliminate the need for 3D annotation.

4.1.1 Image VAE

To leverage the power of transfer learning, a pre-trained ResNet-18 [9] with two additional linear layers one for mean and another for log-variance is used as the encoder and a series of five 2D convolutional layers, each followed by a batch normalization and an activation function like ReLU or Tahn is used as the image decoder.

4.1.2 Pose VAE

For the sake of simplicity and consistency with the previous works for benchmarking the performance, we use a series of 5 linear and ReLU activation blocks with additional linear layers for mean and log-variance for the 2D pose encoder and a linear later for upsampling means to hidden dimensions of the main linear blocks.

4.2 Training Scheme and Loss Function

Training a VAE is a notoriously difficult task, as it involves optimizing not just the reconstruction loss but also the Kullback–Leibler Divergence (KLD) loss. With crossmodal training the number of metrics to optimize increases multi-fold. As described in section[2.1.4], The training scheme for crossmodal training (for crossmodal generation) involves training combinations of encoder and decoder of either the same or different modalities in the same epoch. The reconstruction loss function of that particular combination depends on the decoder. The image decoder uses Least Absolute Deviations (L1) loss and the 3D pose decoder uses Mean Squared Error (MSE) loss. Though the KLD loss is the same for both, it is normalized with the number of elements in the reconstruction, i.e 16*3 for 3D pose and 256*256*3 for RGB images.

4.3 Evaluation Metrics

3D human pose and Human3.6M in particular is mainly evaluated by Mean Per Joint Position Estimate (MPJPE) metric. MPJPE as it literally abbreviates, is the mean of the position estimate for all the joints of a pose. Where per-joint position estimate is nothing but the euclidian distance (usually measured in mm) between the predicted joint to its ground truth.

Chapter 5

Results

5.1 Human3.6M

5.1.1 Evaluation protocol

Human3.6M has 11 subjects out of which 7 are publically released while the rest are kept private. There are 2 widely used evaluation protocol. Protocol-1 is using all 4 camera views in subjects $S1$, $S5$, $S6$, $S7$ and $S8$ for training and the same 4 camera views in subjects $S9$ and $S11$ for validation/testing. Protocol-2 is the same as 1 expect that the predictions are post-processed via a rigid transformation before comparing to the ground-truth.

5.1.2 2D-3D lifting

The current experiments and results are only for 2D to 3D lifting VAE and could be found at https://app.wandb.ai/b-sridatta/hpe3d?workspace=user-b-sridatta. The results illustrated below are after training the model for 100 epochs on around 400,000 2D poses without augmentation. The architecure is as described in earlier with 512 hidden units per linear layer and 100 latent dimension and with a beta weight for KLD loss as 0.001. Further immediate experiments that have to be carried out are itegrating data augmentation, beta annealing, lower latent dimensions and image modalities.

Figure [5.1.1] illustrates 25 random predictions of the validation poses (in blue) and their correspsonding error (in red) w.r.t the ground truth (in gray) in millimeters.

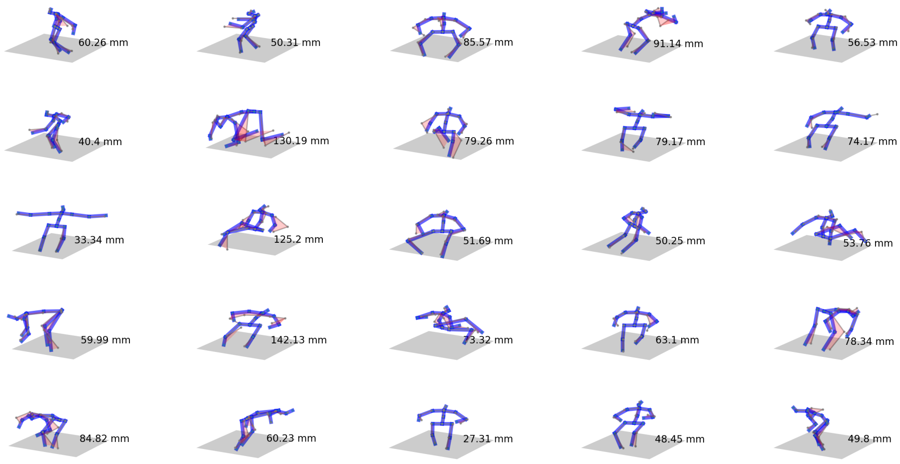

Figure 5.1.1: Comparing predictions and ground truth

Figure[5.1.2] is the visualization of 2D pose embedding in latent space after dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) with different number of neighbours where small neighbours extract local features and large numbers extract more global features. Each action is given a unqiue color. Though we small clusters of blues, browns, pinks the overall space looks very mixed up. This is expected to be improved after annealing beta from 0 to 1 over t he course of training or by using cyclical annealing [6]. However, it could also be the case that the many of the instance in different actions over lap. For example, the action standing up and sitting down have instances while both or standing or sitting etc. This could be verified by visualising the latent space with images rather than just points.

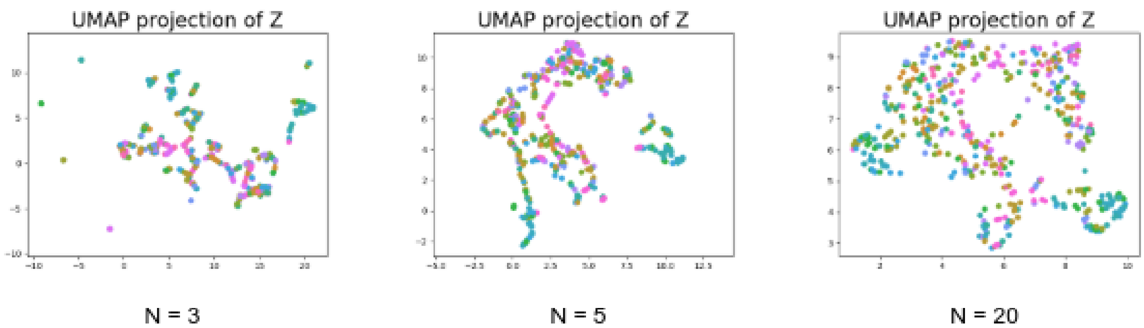

Figure 5.1.2: UMAP Visualization of samples in latent space with varying nearest neigbours

Bibliography

[1] Chang, Ju Yong, Moon, Gyeongsik, and Lee, Kyoung Mu. "AbsPoseLifter: Absolute 3D Human Pose Lifting Network from a Single Noisy 2D Human Pose". In: *arXiv preprint arXiv:1910.12029* (2019).

[2] Chen, Ching-Hang, Tyagi, Ambrish, Agrawal, Amit, Drover, Dylan, Rohith, M. V., Stojanov, Stefan, and Rehg, James M. "Unsupervised 3D Pose Estimation with Geometric Self-Supervision". In: *CoRR* abs/1904.04812 (2019). arXiv: `1904.04812`. URL: `http://arxiv.org/abs/1904.04812`.

[3] Cheng, Yu, Yang, Bo, Wang, Bo, Wending, Yan, and Tan, Robby T. "Occlusion-Aware Networks for 3D Human Pose Estimation in Video". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 723–732. DOI: `10.1109/ICCV.2019.00081`. URL: `https://doi.org/10.1109/ICCV.2019.00081`.

[4] Dario Amodei, Danny Hernandez. *AI and Compute*. `https://openai.com/blog/ai-and-compute/`. (Accessed on 05/17/2020).

[5] Fabbri, Matteo, Lanzi, Fabio, Calderara, Simone, Palazzi, Andrea, Vezzani, Roberto, and Cucchiara, Rita. "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World". In: *European Conference on Computer Vision (ECCV)*. 2018.

[6] Fu, Hao. "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing". In: *NAACL*. 2019.

[7] Gu, Jiajun, Wang, Zhiyong, Ouyang, Wanli, Zhang, Weichen, Li, Jiafeng, and Zhuo, Li. "3D Hand Pose Estimation with Disentangled Cross-Modal Latent Space". In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 2020,

pp. 380–389. DOI: 10.1109/WACV45572.2020.9093316. URL: https://doi.org/10.1109/WACV45572.2020.9093316.

[8] Hanbyul Joo, Hao Liu. “Panoptic Studio: A Massively Multiview System for Social Motion Capture”. In: (2015).

[9] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[10] *HumanEva : Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion | SpringerLink*. https://link.springer.com/article/10.1007/s11263-009-0273-6. (Accessed on 06/08/2020).

[11] Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339.

[12] Kong, Chen and Lucey, Simon. “Deep Non-Rigid Structure From Motion”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1558–1567. DOI: 10.1109/ICCV.2019.00164. URL: https://doi.org/10.1109/ICCV.2019.00164.

[13] Kudo, Yasunori, Ogaki, Keisuke, Matsui, Yusuke, and Odagiri, Yuri. “Unsupervised Adversarial Learning of 3D Human Pose from 2D Joint Locations”. In: *CoRR* abs/1803.08244 (2018). arXiv: 1803.08244. URL: http://arxiv.org/abs/1803.08244.

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

[15] Martinez, Julieta, Hossain, Rayat, Romero, Javier, and Little, James J. “A Simple Yet Effective Baseline for 3d Human Pose Estimation”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2659–2668. DOI: 10.1109/ICCV.2017.288. URL: https://doi.org/10.1109/ICCV.2017.288.

[16] Moon, Gyeongsik, Chang, Ju Yong, and Lee, Kyoung Mu. "Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image". In: *CoRR* abs/1907.11346 (2019). arXiv: `1907.11346`. URL: `http://arxiv.org/abs/1907.11346`.

[17] Novotny, David, Ravi, Nikhila, Graham, Benjamin, Neverova, Natalia, and Vedaldi, Andrea. "C3DPO: Canonical 3d pose networks for non-rigid structure from motion". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 7688–7697.

[18] Shi, Yuge, Siddharth, N., Paige, Brooks, and Torr, Philip H. S. "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". In: *CoRR* abs/1911.03393 (2019). arXiv: `1911.03393`. URL: `http://arxiv.org/abs/1911.03393`.

[19] Spurr, Adrian, Song, Jie, Park, Seonwook, and Hilliges, Otmar. "Cross-Modal Deep Variational Hand Pose Estimation". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 89–98. DOI: `10.1109/CVPR.2018.00017`. URL: `http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Spurr%5C_Cross-Modal%5C_Deep%5C_Variational%5C_CVPR%5C_2018%5C_paper.html`.

[20] Wan, Chengde, Probst, Thomas, Gool, Luc Van, and Yao, Angela. "Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1196–1205. DOI: `10.1109/CVPR.2017.132`. URL: `https://doi.org/10.1109/CVPR.2017.132`.

[21] Wandt, Bastian and Rosenhahn, Bodo. "RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation". In: *CoRR* abs/1902.09868 (2019). arXiv: `1902.09868`. URL: `http://arxiv.org/abs/1902.09868`.

[22] Wang, Chaoyang, Kong, Chen, and Lucey, Simon. "Distill Knowledge From NRSfM for Weakly Supervised 3D Pose Learning". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea*

(South), October 27 - November 2, 2019. IEEE, 2019, pp. 743–752. DOI: 10.1109/ICCV.2019.00083. URL: https://doi.org/10.1109/ICCV.2019.00083.

[23] Wang, Chaoyang, Lin, Chen-Hsuan, and Lucey, Simon. "Deep NRSfM++: Towards 3D Reconstruction in the Wild". In: *CoRR* abs/2001.10090 (2020). arXiv: 2001.10090. URL: https://arxiv.org/abs/2001.10090.