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

Unsupervised 3D Human Pose Estimation

KTH Thesis Report
Draft for final thesis meeting

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

Abstract

YET TO BE WRITTEN

This is a template for writing thesis reports for the ICT school at KTH. I do not own any of the images provided in the template and this can only be used to submit thesis work for KTH.

The report needs to be compiled using XeLaTeX as different fonts are needed for the project to look like the original report. You might have to change this manually in overleaf.

This template was created by Hannes Rabo <hannes.rabo@gmail.com or hrabo@kth.se> from the template provided by KTH. You can send me an email if you need help in making it work for you.

Write an abstract. Introduce the subject area for the project and describe the problems that are solved and described in the thesis. Present how the problems have been solved, methods used and present results for the project. Use probably one sentence for each chapter in the final report.

The presentation of the results should be the main part of the abstract. Use about ½ A4-page. English abstract

Keywords

Template, Thesis, Keywords ...

Abstract

YET TO BE WRITTEN

Svenskt abstract Svensk version av abstract – samma titel på svenska som på engelska.

Skriv samma abstract på svenska. Introducera ämnet för projektet och beskriv problemen som löses i materialet. Presentera

Nyckelord

Kandidat examensarbete, ...

Acknowledgements

YET TO BE WRITTEN

Write a short acknowledgements. Don't forget to give some credit to the examiner and supervisor.

Acronyms

AR/VR Augmented Reality/Virtual Reality
GAN Generative Adversarial Network
HPE Human Pose Estimation
KLD Kullback–Leibler Divergence
MPJPE Mean Per Joint Position Estimate
PJPE Per Joint Position Estimate
SOTA State-of-The-Art
UMAP Uniform Manifold Approximation and Projection
VAE Variational Auto-Encoder

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 [2]. 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 Human Pose Estimation (HPE), is a fundamental computer vision task that also forms a basis for advanced tasks such as human action and gesture recognition as well as human motion prediction. 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 the focus of this thesis is on 3D HPE.

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 research explores various ways to solve the task using RGB/Depth image channels, 2D poses, 3D poses, multi-view, and sequential images/videos.

Most of them are supervised learning approaches that require 3D ground truth poses that can only be acquired using physical sensors. Supervised learning methods that learn 3D pose from images, follow a complex cascading approach with 2D poses in some form is an intermediate output. And other learning approaches mostly make use of images from multiple views to estimate the 3D pose.

Assuming 2D poses are obtained from the State-of-The-Art (SOTA) models specialized on 2D HPE models, some works focus on estimating the 3D pose from these 2D poses instead of images. Such networks are called *lifting* networks. These lifting networks can be simple without requiring computationally expensive convolutional layers as it only needs to learn the features of 2D poses that are low dimensional compared to images. However since 2D poses are naturally obtained by projecting 3D poses to a plane, it is an inverse problem. Also, there are multiple feasible 3D poses that when the projected result in the same 2D pose. Thus making the task of lifting 2D-to-3D, a *ill-posed inverse* problem due to its inherent ambiguity.

Non-supervised (Weakly/Self/Unsupervised) learning regimes, that are less dependent on 3D pose ground truth, have also gained traction in recent years. Weakly supervised approaches use 3D ground truth indirectly by generating more 2D poses from more views or, for training a discriminator network of a Generative Adversarial Network (GAN). While unsupervised learning (self-supervised) approaches do not use 3D ground truth poses in any shape or form. Many of the deep learning techniques that have already improved the results in other computer vision tasks are yet to be explored in 3D HPE.

1.2 Problem

How can we learn a strong visual representation of the data to tackle the task of 3D HPE? 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 unsupervised learning approaches and 2D-to-3D lifting methods that could help tackle the challenges in scaling 3D HPE to the

real-world.

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, behavior 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 also plays a vital 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 a regular linear neural network or using various other architectures like autoencoders. These models can use linear, convolutional, and graph layers to learn features. This thesis focuses on investigating the merit in using architectures like Variational Auto-Encoders (VAEs) to solve the 3D HPE within the context of leveraging probabilistic inference models, 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 3D pose thus requiring 3D annotations, or by projecting the prediction back to 2D to compare with the input

(requires only 2D annotations that could be acquired from SOTA in 2D HPE) and use a different technique to ensure the correctness of pose in 3D. Adversarial training and self-supervision techniques have also given promising results in the last couple of years. The method proposed in the thesis is designed to learn 3D from 2D poses alone in an unsupervised-adversarial learning fashion after the capabilities of the method under supervised settings being verified.

The prime motivation behind the design choices is to address the challenges in scaling up 3D HPE to real-world.

1.6 Stakeholders

Daimler's 'Environment Perception for Autonomous Driving' R&D team in Stuttgart, Germany, 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 understanding the human state, motion, and behavior, 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 HPE is a critical element to accomplish this 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 contributions 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 and uses only publicly available, widely used, and benchmarked datasets.

1.8 Outline

The theoretical knowledge required to understand the details of the thesis is presented next in chapter **??**, Theoretical Background. This entails the explanation of some preliminary concepts **??**, research area introduction **??**, where the vast literature related to HPE is summarized, and the highlights of all the works **??** that are closely related to the thesis.

The background is followed by chapter **??**, Data, providing details of the datasets used along with some visualizations. This chapter also explains the 3D projective geometry concepts required to understand the pre and post-processing steps the data undergoes.

The method, chapter **??**, describes the proposed approach in detail. This covers the components of the proposed architecture, the motivation behind the choices, the training and validation procedure, and other details that help reproduction.

The results are analysed and discussed in chapter 2 and conclusions are presented in chapter 3.

Chapter 2

Results

2.1 Quantitative Results

The results presented here are after training the networks for $\sim$400 epochs (5.5 hours) on approximately 300,000 2D poses with a batch size of 2560 on an Nvidia Titan X. The input poses are flipped with a probability of 0.5. The model takes 16 joints as the output where the root is added at the origin for validation. The proposed architecture consists 1024 hidden units per linear layer and 51 latent dimensions. Both the Variational Auto-Encoder (VAE) and the discriminator are trained using Adam optimizer with default hyperparameters and with a learning rate of 2e-4. The gradient norms of the discriminator is clipped to 1 when training the discriminator. While training the generator the gradient norms are clipped to 2 for all the models while the gradient values are clipped to 1000.

One of the challenging parts is finding the optimal weights for each of the terms in the triplet loss. The loss coefficients λ_{recon}, λ_{KLD}, $\lambda_{disc.}$ are set to 1, 0.001, 0.001 respectively. The higher weight is motivated by 2 reasons. λ_{recon} refers to the constrained optimization and irrespective of how realistic it is, projection loss is desired to be consistently low to get better Mean Per Joint Position Estimate (MPJPE). That leads to the other reason that the quantitative results are given higher importance.

The values of λKLD and $\lambda disc.$ can be tuned according to the task at hand based on how well the poses are to be clustered or how important it is to reject poses that are not realistic. The β value for the VAE is cycled from 0 to λ_{KLD} every 40 epochs. While

keeping it constant at λ_{KLD} for 10 epochs with a 10 epoch warmup at the beginning of the training.

The results obtained by the networks with the above configuration in addition to the choices mentioned in **??** are summarized in Table 2.1.1.

Supervision	Algorithm	Error (mm)
Full	Martinez *et al.* [6]	37.1
	Chen *et al.* [4] (SH, MH)	42.6
Weak	3D Interpreter *et al.* [9]	88.6
	AIGN *et al.* [7]	79.0
	Wandt *et al.* [8]	38.2
	Drover *et al.* [3]	38.2
	Chen *et al.* [5] (SH)	48.7
	Chen *et al.* [5] (BH)	31.6
Unsupervised	Ching *et al.* [1]	58
	Ching *et al.* [1] (AD) (TD)	51
	Ours	52.4
	Ours (BH)	50

Table 2.1.1

The different transformations of the poses during the training are presented in the proposed architecture illustration Fig. **??**.

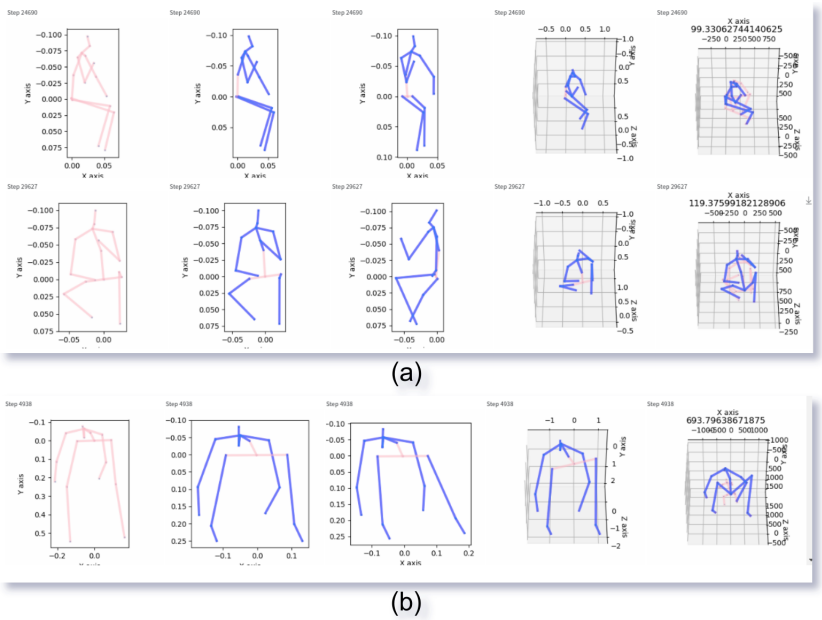

Figure 2.1.1: (a) Prediction on hard poses with hih ambiguity. (b) Poses that can be improved with changes to data processing.

Despite the MPJPE being ~ 68, the decoder generates many accurate 3D poses that

are almost indistinguishable for the human eye. The high number of outliers and lower performance on hard poses worsens the overall performance of the model. The graph on the left depicted in fig 2.1.2 shows the slow and gradual decrease in MPJPE. And the graph on the right shows the histogram of the Per Joint Position Estimate (PJPE) of each sample over time. This shows the model is unable to certain poses while it improves gradually on the rest.

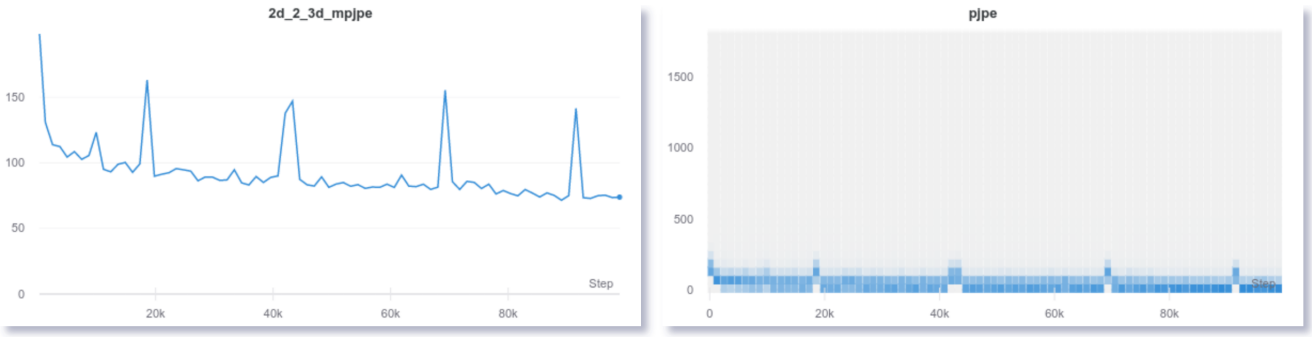

Figure 2.1.2: MPJPE and PJPE trends during the training respectively

Some of such outliers are presented in fig 2.1.1, the predictions in (a) are the ones the model is unable to learn. While (b) is the evidence of the shortcoming of the current processing technique. Rectifying that would improve the evaluation metric of the model quite significantly.

The visualization of 2D pose embedding in latent space after dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) is shown in fig. 2.1.3. Each action is given a unique color. Though we small clusters of blues, browns, pinks the overall space looks very mixed up. This is expected as many of the instances in different actions overlap. For example, the action standing up and sitting down have instances while both or standing or sitting, etc.

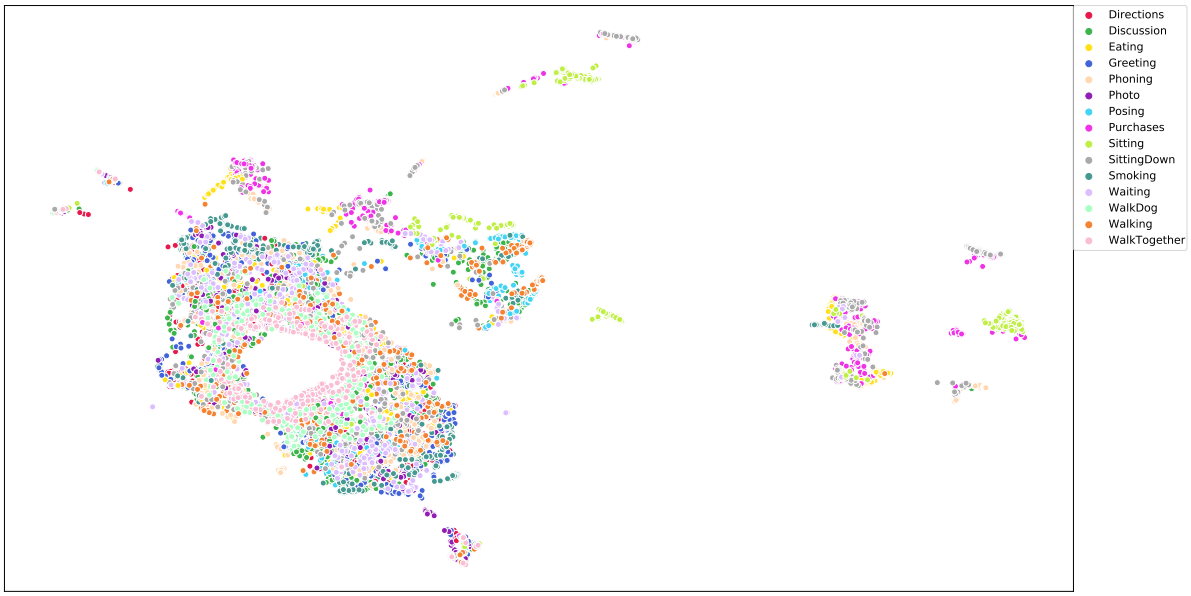

Figure 2.1.3: UMAP Visualization of samples in latent space. The actions do not always directly related to the pose due to overlaps from one action to another. Better viewed in color and 200% zoom

Chapter 3

Conclusions

YET TO BE WRITTEN

Describe the conclusions (reflect on the whole introduction given in Chapter 1).

Discuss the positive effects and the drawbacks.

Describe the evaluation of the results of the degree project.

Describe valid future work.

The sections below are optional but could be added here.

3.1 Discussion

3.1.1 Future Work

3.1.2 Final Words

Bibliography

[1] 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.

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

[3] Drover, Dylan, MV, Rohith, Chen, Ching-Hang, Agrawal, Amit, Tyagi, Ambrish, and Huynh, Cong Phuoc. *Can 3D Pose be Learned from 2D Projections Alone?* 2018. arXiv: 1808.07182 [cs.CV].

[4] Li, Chen and Lee, Gim Hee. *Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network*. 2019. arXiv: 1904.05547 [cs.CV].

[5] Li, Chen and Lee, Gim Hee. *Weakly Supervised Generative Network for Multiple 3D Human Pose Hypotheses*. 2020. arXiv: 2008.05770 [cs.CV].

[6] 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.

[7] Tung, Hsiao-Yu Fish, Harley, Adam W., Seto, William, and Fragkiadaki, Katerina. *Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision*. 2017. arXiv: 1705.11166 [cs.CV].

[8] 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.

[9] Wu, Jiajun, Xue, Tianfan, Lim, Joseph J., Tian, Yuandong, Tenenbaum, Joshua B., Torralba, Antonio, and Freeman, William T. “Single Image 3D Interpreter Network”. In: *Lecture Notes in Computer Science* (2016), pp. 365–382. ISSN: 1611-3349. DOI: 10.1007/978-3-319-46466-4_22. URL: http://dx.doi.org/10.1007/978-3-319-46466-4_22.

Appendix - Contents

Appendix A

Model Summaries

A.1 Encoder

	Layers	Parameters
Upsampling Block		
	Linear(in features=32, out features=1024, bias=True)	33792
	BatchNorm1d(1024, eps=1e-05, momentum=0.1)	2048
	Mish()	0
	Dropout(p=0.2, inplace=False)	0
Residual Block		
	Linear(in features=1024, out features=1024, bias=False)	1048576
	BatchNorm1d(1024, eps=1e-05, momentum=0.1)	2048
	Mish()	0
	Dropout(p=0.2, inplace=False)	0
	Linear(in features=1024, out features=1024, bias=False)	1048576
	BatchNorm1d(1024, eps=1e-05, momentum=0.1)	2048
	Mish()	0
	Dropout(p=0.2, inplace=False)	0
Downsampling Block		
	Linear(in features=1024, out features=51, bias=True)	52275
	Linear(in features=1024, out features=51, bias=True)	52275

A.2 Decoder

	Layers	Parameters
Upsampling Block		
	Linear(in features=51, out features=1024, bias=True)	53248
	BatchNorm1d(1024, eps=1e-05, momentum=0.1, affine=True, track running stats=True)	2048
	Mish()	0
	Dropout(p=0.2, inplace=False)	0
Residual Block		
	Linear(in features=1024, out features=1024, bias=False)	1048576
	BatchNorm1d(1024, eps=1e-05, momentum=0.1, affine=True, track running stats=True)	2048
	Mish()	0
	Dropout(p=0.2, inplace=False)	0
	Linear(in features=1024, out features=1024, bias=False)	1048576
	BatchNorm1d(1024, eps=1e-05, momentum=0.1, affine=True, track running stats=True)	2048
	Mish()	0
	Dropout(p=0.2, inplace=False)	0
	Linear(in features=1024, out features=48, bias=True)	49200
Downsampling Block		
	Linear(in features=1024, out features=48, bias=True)	49200
	Tanh()	0

A.3 Discriminator

	Layers	Parameters
Upsampling Block		
	Linear(in features=32, out features=1024, bias=True)	33792
	LeakyReLU(negative slope=0.01)	0
Residual Block		
	Linear(in features=1024, out features=1024, bias=True)	1049600
	LeakyReLU(negative slope=0.01)	0
	Dropout(p=0.5, inplace=False)	0
	Linear(in features=1024, out features=1024, bias=True)	1049600
	LeakyReLU(negative slope=0.01)	0
	Dropout(p=0.5, inplace=False)	0
Residual Block		
	Linear(in features=1024, out features=1024, bias=True)	1049600
	LeakyReLU(negative slope=0.01)	0
	Dropout(p=0.5, inplace=False)	0
	Linear(in features=1024, out features=1024, bias=True)	1049600
	LeakyReLU(negative slope=0.01)	0
	Dropout(p=0.5, inplace=False)	0
Upsampling Block		
	Linear(in features=1024, out features=1, bias=True)	1025
	Sigmoid()	0