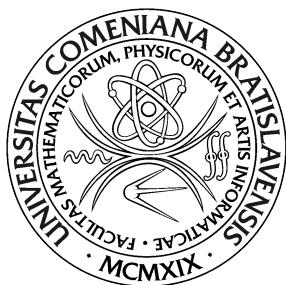


COMENIUS UNIVERSITY BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND
INFORMATICS



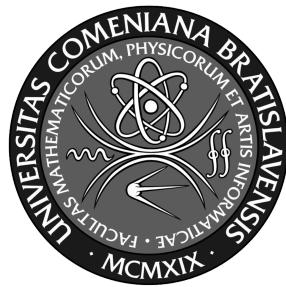
SKELETON TRACKING USING DEEP LEARNING

Master thesis

2020

Bc. Dana Škorvánková

COMENIUS UNIVERSITY BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND
INFORMATICS



SKELETON TRACKING USING DEEP LEARNING

Master thesis

Program of Study: Applied Informatics
Field of Study: 2511 Applied Informatics
Department: Department of Applied Informatics
Supervisor: RNDr. Martin Madaras, PhD.

Bratislava, 2020

Bc. Dana Škorvánková



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Dana Škorvánková

Študijný program: aplikovaná informatika (Jednooborové štúdium,
magisterský II. st., denná forma)

Študijný odbor: aplikovaná informatika

Typ záverečnej práce: diplomová

Jazyk záverečnej práce: anglický

Sekundárny jazyk: slovenský

Názov: Skeleton Tracking using Deep Learning
Skeleton Tracking using Deep Learning

Anotácia: (C++, neural network)

Ciel:
- skeleton tracking of a person
- input is a sequence of RGB or RGBD images
- use neural network to approximate the body movement

Vedúci: RNDr. Martin Madaras, PhD.

Katedra: FMFI.KAI - Katedra aplikovej informatiky

Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 25.09.2017

Dátum schválenia: 03.10.2018

prof. RNDr. Roman Ďuríkovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

I hereby declare I wrote this thesis by myself, only
with the help of referenced literature, under the careful
supervision of my thesis supervisor.

.....

Bratislava, 2020

Bc. Dana Škorvánková

Acknowledgement

I would like to thank ...

Abstrakt

Cieľom tejto diplomovej práce je vyvinúť systém na 3D odhad pózy človeka, založený na neurónovej sieti, ktorý má na vstupe trojdimenzionálne dátá (vo forme mračna bodov alebo hĺbkovej mapy) a na výstupe vracia 3D koordináty vrcholov kostry. Naším zámerom je predstavenie nášho vlastného prístupu, ako aj implementácia niekoľkých existujúcich modelov, vynikajúcich svojimi výsledkami v rámci danej problematiky. Jednotlivé metódy následne evaluujeme na referenčných databázach, a porovnáme s doposiaľ najlepšími výsledkami dosiahnutými na uvedených dátach.

Kľúčové slová: 3D odhad pózy človeka, hlboké učenie, neurónové siete, mračná bodov

Abstract

The aim of our thesis is to develop a 3D human pose estimation pipeline based on neural network, which takes three-dimensional data (in a form of a point cloud or a depth map) as input and outputs the 3D skeletal joint coordinates. Our goal is to introduce our own approach, in addition to implementing several well-performing models proposed in existing papers. Next, we aim to evaluate the methods on multiple benchmark datasets, and compare the results to the current state-of-the-art.

Keywords: 3D human pose estimation, deep learning, neural networks, point clouds

Contents

1	Introduction	1
2	Motivation	4
3	Overview	7
3.1	Neural networks	7
3.1.1	Image processing using deep learning	8
3.1.2	3D pose estimation	9
4	Related work	11
4.1	Human pose estimation from RGB data	12
4.2	Depth-based human pose estimation	14
4.3	Point cloud input data	16
5	Proposed implementation	20
5.1	Technologies	20
5.2	Benchmark datasets	20
5.2.1	ITOP	21
5.2.2	UBC3V	22
5.2.3	MHAD	23
5.2.4	CMU Panoptic Dataset	23

<i>CONTENTS</i>	ix
5.3 Approaches	25
5.3.1 Existing models	25
5.3.2 Segmentation-Guided Pose Estimation	27
6 Implementation	29
6.1 Re-implementation of existing models	29
6.1.1 Deep Depth Pose model	29
6.1.2 Point-Based Pose Estimation model	34
6.2 Segmentation-Guided Pose Estimation	42
7 Results	47
7.1 Evaluation metrics	47
7.2 Experiments	47
8 Conclusion	54

List of Tables

7.1	Mean per joint position error on MHAD dataset, compared to state-of-the-art methods.	50
7.2	Mean per joint position error on UBC3V hard-pose dataset, compared to state-of-the-art methods.	52

List of Figures

4.1	The overview of the VNect pipeline.	12
4.2	The V2V-PoseNet architecture.	15
4.3	The PointNet++ hierarchical feature learning architecture.	17
5.1	Sample depth images from the ITOP dataset.	21
5.2	Sample data from the UBC3V dataset.	22
6.1	Deep Depth Pose model architecture.	30
6.2	Our modification of the original DDP model architecture.	33
6.3	The Point-Based Pose Estimation model architecture.	35
6.4	Data pre-processing shown on a sample point cloud from MHAD dataset.	39
6.5	The skeleton structure used in MHAD dataset before and after modification.	40
6.6	The automatic body region annotation on real data.	41
6.7	The Segmentation-Guided Pose Estimation pipeline.	44
7.1	Mean average error per joint on ITOP and CMU dataset.	48
7.2	Mean average error per joint on MHAD and UBC3V dataset.	48
7.3	Qualitative results of our method on CMU Panoptic dataset.	49
7.4	Mean average precision at threshold on MHAD dataset.	51

7.5	Qualitative results on test set of UBC hard-pose dataset.	52
7.6	Accuracy of the body-parts segmentation of our method.	53

Chapter 1

Introduction

Deep learning was first introduced as a machine learning research area in 1986, with the aim of shifting the concept of machine learning closer towards artificial intelligence. During the last decade, artificial neural networks have become one of the most frequently used methods of machine learning in various research fields, such as computer vision, robotics, medicine, manufacturing, telecommunications, automotive engineering, and many more.

Neural networks are able to carry out numerous different tasks, which essentially consist of classification, prediction, clustering and associating. Regarding the classification task, the neural network learns to organize patterns or data into a number of predefined categories. Classification algorithms are often used to solve issues like medical diagnoses, e-mail spam filtering, speech recognition, handwriting recognition or image recognition. Concerning prediction (or regression) tasks, the aim is to produce the expected output from the given input data. Clustering is an unsupervised task, which concludes of a classification of input data based on an identification of a unique feature of the data, without any predefined classes. This technique is widely used

for pattern recognition, feature extraction, data mining, image segmentation etc. The associating task means the neural network is capable of storing or remembering certain patterns, thus it associates the previously unseen data with the most comparable pattern in its memory. This is mostly used in the field of pattern recognition and pattern completion.

One of many fields where the neural networks are applicable is human motion analysis. Some of the most frequent motion tasks include skeleton tracking, human motion prediction and pose estimation. The motion tasks using either data-based or physics-based methods still remain a challenge these days. The data-driven methods rely mostly on motion capture systems, while the physics-based methods depend on optimization to predict motion.

In our thesis, we will be focusing on the task of pose estimation. The main goal of our study is to implement a method for 3D human pose estimation from depth input data using a deep learning approach. Our intention is to build a convolutional neural network to perform the prediction of 3D skeletal joint positions. In our work, we are going to implement several models and evaluate the results obtained on a number of benchmark datasets.

The thesis is organized in the following manner: In the next chapter of our work, we are going to illustrate basic motivation that lead us to devote ourselves to the stated issue, and indicate the current deficiencies in the problematics. In the third chapter, we will summarize an overview of neural networks used in image processing and take a look specifically on the 3D pose estimation task. The current state of research in this field and related work is discussed in the fourth chapter. The fifth chapter is dedicated to our proposed implementation of the given task, that is, the technologies we

are going to use for the implementation, the benchmark datasets on which we will evaluate our results, and the models we are going to implement. In the subsequent chapter, we will focus on the actual implementation of the proposed methods, we will go through implementation details, training procedures, and problems encountered during the process. The evaluated results of our thesis are reviewed in the seventh chapter. Finally, in the last chapter, we will sum up the goals of our thesis, the achieved results, and the conclusions of our work.

Chapter 2

Motivation

The task of human pose estimation attracts a lot of attention among deep learning researchers, mainly because of its frequent usage in virtual and augmented reality, action recognition, ergonomic body posture analysis, surveillance, human-robot interaction, trajectory prediction or motion-based human identification. Although a lot has been achieved in the 3D human pose estimation task, there are still many challenges nowadays, which are not easy to overcome.

Since most of the research is currently focused on estimating the pose from RGB data, one of the most critical challenges of pose estimation from 3D input is data availability. To successfully train a neural network of reasonable size, a large and well labeled dataset is crucial. Right now, there is a very small set of publicly available 3D human pose estimation databases. Moreover, even among the available datasets, it is hard to find one that is both large enough in its scale, and accurate enough to avoid overfitting of the neural model. There are several large action recognition datasets with motion capture ground truth, but since providing the exact skeleton joint

locations is not their primal purpose, the ground truth is often not accurate enough for the task of pose estimation.

Due to the lack of the accessible depth data, many researchers have recently used their own recorded depth datasets to evaluate the results of their proposed method. However, this leads to the fact, that it is difficult to objectively compare the particular methods to each other, because the recorded databases are often not published. It is important to mention, that recording of a quality depth dataset is not a trivial task, mainly since the expensive motion capture system is usually required to obtain accurate ground truth labels, which also limits us to indoor scenes. The limited accuracy of the ground truth poses is usually caused by poor synchronization of a depth sensor and a motion capture system. The most commonly used depth sensors do not have a stable frame rate, which results in time delays and misalignment between frames, and makes the precise synchronization practically impossible. In some of the datasets, this issue is partly fixed by time-stamping technique, refining the frame alignment, and filtering out the mismatches. It is even harder considering the multi-view approach, when the multiple depth sensors need to be synchronized mutually as well as with the motion capture system.

Another issue concerning pose estimation from 3D data is the actual type of 3D data that is passed as input to the neural network. The most frequent option is to use depth maps, thus encoding the third dimension into the 2D image. The depth maps are a very dense representation of a human pose, which results in expensive computations and lowering the time efficiency, while also processing the seemingly redundant data. Furthermore, since depth maps are usually treated by neural networks as 2D images, there

arises the same problem as in estimating 3D pose from RGB data, i.e. the need for highly non-linear operations. Additionally, because of the projection of an object in 3D space onto a 2D image plane, the actual shape of the human pose can be distorted in the depth map, which means the network has to perform the perspective distortion-invariant estimation. In an attempt to overcome these drawbacks, voxelized grids have been used in several solutions to provide sparser 3D data representation. Despite that, voxels have their shortcomings, too. First of all, voxels require 3D convolution operations, which are rather demanding in terms of memory, time, and computing power. Moreover, the conversion of point clouds or depth maps into the voxelized grids can be time-consuming itself.

Chapter 3

Overview

In this chapter, we will go through general overview of the techniques we are going to use in our solution. We will define basic terms associated with neural networks and their functioning, illustrate several scenarios on how neural networks can be utilized in image processing, and explain the concept of the pose estimation task, along with the various sub-categories we distinguish among.

3.1 Neural networks

In general, artificial neural networks were introduced as structures inspired by the biological structure of human brain. Similarly to the brain, the basic computational units of the artificial neural networks are neurons, which are connected together by synapses. In neural networks, synapses are simply weight values, which means the neuron performs some kind of calculation and the result is multiplied by the value corresponding to the particular connection it is passing through. The neurons are organized in layers, which is the main idea of the so-called deep learning. The neural networks can

perform both supervised and unsupervised learning.

In most cases, we are dealing with the supervised learning, where we feed the model with the input data and also provide the output we expect the network to produce. As the input data is passed through the network in what we call a forward-pass, the neurons are outputting certain computed values and passing them to the neurons they are connected to, which in the end form a single or multiple outputs of the network. The idea of the supervised learning is that at the end of the forward-pass, we provide a feedback to the network, about how correctly it performed the desired task. This is done by comparing the actual output of the model to the desired output. The comparison is also referred to as calculating a loss function. The calculated loss is then back-propagated through all the layers in the network to adjust the weights of the model accordingly. This scenario is repeated in many iterations, while the loss value computed on the model output should be decreasing.

Nowadays, as we have enough computational power and the amount of available data large enough to train on, the capacities and architectures of the neural networks are becoming increasingly extensive and powerful.

3.1.1 Image processing using deep learning

Neural networks have shown their significant contribution to various research fields, and one of them is digital image processing. Many different tasks concerning the image processing can be carried out by a neural network, e.g. image classification, feature extraction, pattern recognition, object detection, image captioning etc. For instance, in image classification and feature extraction, deep learning has shown many advantages over the traditional

methods, which rely on handcrafted features. Using a neural network model, instead of manually selecting and extracting image features, we can pass the image directly and the model learns to find and distinguish the important features itself, and additionally, might classify the image accordingly. While the first layers often recognize very simple features, like edges and corners, the deeper we dive into the network architecture, the more complex features are sought in particular layers and neurons within them.

3.1.2 3D pose estimation

One of the tasks concerning image processing and human motion analysis, frequently performed by neural networks, is the 3D pose estimation. As the name implies, it is a task of estimating the three-dimensional pose of a human subject from a single image or set of video frames. The resulting pose is determined by 3D coordinates of the skeletal joints of the human body. The number of joints can vary, thus it is usually considered a hyperparameter. The more skeletal joints (and output coordinates) the model estimates, the more complex representation of the human body skeleton we obtain.

The human pose estimation approaches can be classified into model-based generative methods and discriminative methods. Generative methods treat the human body as an articulated structure. The utilized model generally describes the appearance of particular body parts and the spatial relationship between adjacent parts. Discriminative methods solve the task of pose estimation by direct regression from feature space to pose space, and are invariant to body shape variations. Deep learning approaches avoid the manual dealing with features and structural dependencies by embedding it into the mapping function and learning high-level representation of the input data.

We mentioned the neural model infers the pose from an image, however, the actual form of the input data may vary. First, the individual approaches to the task can be divided in two categories according to the input data dimension. Two-dimensional input data are usually RGB images. The important advantage of such approaches is that they can be easily utilized in real-time in-the-wild applications, since a common RGB camera is sufficient to capture the images. This is also a reason, why a majority of the pose estimation research nowadays is focused on inferring from 2D data, as well as most of the public pose estimation datasets consist only of RGB images. On the other hand, three-dimensional input data bring an additional depth information into the neural networks, which is often very handful in increasing the accuracy of the estimation. The three-dimensional input data comes in various forms. The most common are depth maps. The basic idea behind depth maps is to encode the third dimension into the two-dimensional image, where at each pixel location, the corresponding pixel value is representing the value of the third coordinate (or depth) at the specified position. Other than depth maps, a convenient way to capture three-dimensional data is to use point clouds. Anyhow, point clouds have several attributes which are not suitable for neural networks. Most importantly, point clouds are unordered and irregular. Some of the researchers resolved this by making use of the voxelized grids, that is, discretized the point clouds into a predefined grid. Alternatively, several approaches evolved on the idea of processing the unorganized point clouds directly using shared multi-layer perceptrons to obtain the features of the point cloud and work further with them.

Chapter 4

Related work

Nowadays, neural networks are widely used in the field of image processing, pattern recognition, human movement analysis and many more. There are numerous types of tasks concerning human movement analysis, where the neural networks proved to be beneficial, e.g. action recognition, action classification, body-movement-based human identification, pose estimation etc.

Focusing on the pose estimation task, there have been many different methods and approaches presented in recent years. Based on the type of the input data, the studies can be divided into approaches inferring from two-dimensional data (RGB images) [3, 16, 17, 19, 22, 27, 31], and three-dimensional data (depth maps, point clouds, voxelized grids etc.) [1, 4, 5, 6, 9, 20, 29, 30, 36]. The two-dimensional approaches are far more usable and easily accessible in real-time applications, being able to run without any special devices, using only the RGB camera. On the other hand, the regression of 3D joint positions from 2D input data requires highly non-linear operations, what can lead to many difficulties in the learning procedure. The three-dimensional

approaches provide the additional depth information, which can significantly simplify the task for the network, and thus improve the estimation accuracy.

4.1 Human pose estimation from RGB data

We can divide studies working with the RGB input data in two main groups based on whether they directly regress the 3D pose coordinates [15, 32] or use the 2D pose to infer the 3D pose [2, 16, 18, 17, 27]. Among those employing the 2D pose, many approaches make use of lifting the estimated 2D pose to 3D [2, 8, 15, 21] by direct regression, database matching etc.

One of the first real-time approaches working with the RGB data was proposed by Mehta et al. [17]. They introduced a system called *Vnect* to obtain real-time full global 3D skeletal pose, combining a pose regressor based on convolutional neural network with kinematic skeleton fitting. They parametrized each 3D skeletal joint by a confidence heatmap and three location maps, one for each axis. The predicted 3D position of each body joint is read out from the location maps, corresponding to x, y, and z coordinate, at the estimated pixel location of the particular keypoint, determined by the heatmap. Figure 4.1 depicts the pipeline introduced in the stated paper.

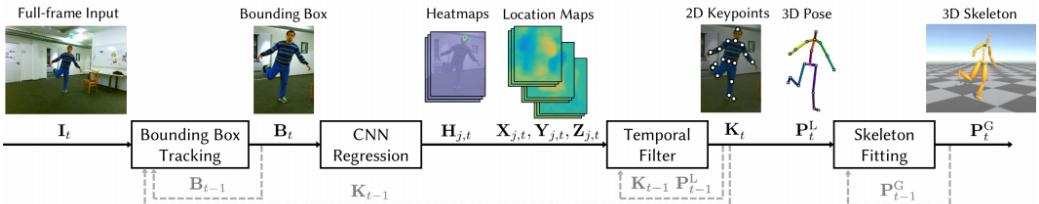


Figure 4.1: The overview of the Vnect pipeline [17].

However, the stated model was unable to handle occlusions or capture multiple people in the scene. The authors removed these restrictions in the follow-up model called *XNect* [19], which is able to capture multiple people in the scene by a single RGB camera. Unlike the previous work, the model outputs full skeletal pose in joint angles and global body positions of a coherent skeleton in real-time. The method consists of three subsequent stages: First, a convolutional neural network estimates 2D and 3D features and identity assignments for all visible body joints of all subjects in the scene, second, a fully-connected neural network turns possibly occluded 2D and 3D pose features into a complete 3D pose per subject, and third, a space-time skeletal model fitting is applied to the predicted 2D and 3D pose to merge them and enforce temporal coherence.

In another paper, Mehta et al. [16] focused on RGB data as well, but this time, they presented a method based on the concept of transfer learning. They embraced the similarity between the feature extraction tasks in 2D and 3D space, and transferred the features learned on 2D data to 3D pose estimation network (that is, pre-trained the 3D convolutional model on 2D features). In the paper, they put a great emphasis on *in-the-wild* scenes, that is, the pipeline was designed to perform well even outside of the recording room, in generic outdoor scenes. Newell [22] presented stacked hourglass networks, consisting of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference. These networks are designed such that the features are processed across all scales, allowing them to capture the spatial relationships within the body. Chou et al. [3] employed generative adversarial networks as the learning paradigm. They consist of two stacked hourglass networks – generator and discriminator, where the former is used for the human pose estimation and the latter back-propagates the adversarial

loss between the ground truth and the generated output to the generator.

Rogez et al. [27] decided to utilize the concept of the pre-defined anchor poses. Their proposed pipeline consists of three sub-tasks. First, they extract candidate regions using a Region Proposal Network [26] and place the set of anchor poses into the predicted bounding boxes. Then, they run it through a classification branch, which outputs probabilities of anchor poses to be correct at each location. Finally, the regression branch computes an anchor-pose-specific regression that estimates the difference between the pose proposal and ground truth pose, and refines the final pose estimation.

4.2 Depth-based human pose estimation

The depth data used as the input to the neural networks comes in various forms. Most frequently used are depth maps [12, 14, 28, 35]. Depth maps are actually encoding 3D space into 2D image, where the value at each pixel position represents the corresponding depth value (third axis coordinate). Marin-Jimenez et al [14] proposed a technique where the final estimated pose is computed as the weighted sum of the predefined set of prototype poses. The weights corresponding to the prototypes are directly regressed from input depth maps by a convolutional neural network. In the paper, they claim to have outperformed state-of-the-art results on the benchmark datasets and reached 100% accuracy in mean average precision at 10cm on the ITOP dataset [6], i.e. every skeleton joint is predicted within 10cm range from its ground truth position. The stated approach is an example of a single-stage method.

The two-stage methods generally consists of the segmentation stage and the

regression stage. First, the input data is segmented to the corresponding body-parts. Then, the segmented input data is used to infer 3D joint coordinates. An example of a two-stage method was proposed by Shafaei et al. [29]. They treat the problem of 3D pose estimation from depth data through a two-stage pipeline, where in the first stage the body parts are identified in the input depth maps by a dense classifier. In the second stage, all camera views are merged, and a set of statistics concerning a created unified 3D point cloud is collected and passed as features to a linear regressor to compute 3D body joint locations.

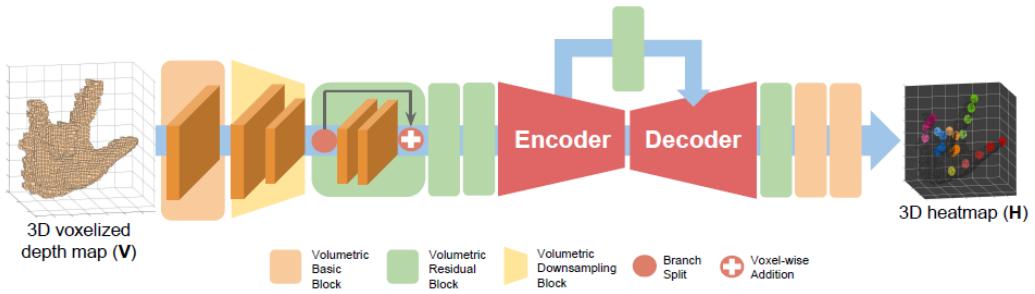


Figure 4.2: The V2V-PoseNet architecture which takes voxelized input and estimates the per-voxel likelihood for each keypoint [20].

Aside from depth maps, some of the methods make use of the voxelized grids, made by discretizing a given point cloud in a predefined set of values [7, 11, 20]. However, voxels require use of three-dimensional convolutions, what makes working with them very time-consuming and computationally expensive. *V2V PoseNet* [20] operates with this kind of data and regresses joint locations with 3D CNN-autoencoders. They first use 3D CNN encoder and decoder to estimate per-voxel likelihood of each skeleton joint from voxelized input (Figure 4.2). Afterwards, they refine the target object localization with a 2D CNN which takes a cropped depth map and output an

offset from its reference point to the center of ground truth joint positions. This way, they obtain an accurate reference point.

4.3 Point cloud input data

As an alternative to depth images or voxels, there are several networks proposed which work directly with unordered point clouds as input data, yet implement the convolution operations on the point clouds without using computationally expensive 3D convolutions. Some of the methods decided to use shared multi-layer perceptrons and max-pooling layers to obtain the features of a point cloud. Although they manage to extract global features, since the max-pooling layers are applied on the whole set of points, it is hard to capture the local context.

Qi et al. [24] proposed a classification and segmentation model called *PointNet*, where they intend to incorporate the local features by an aggregation of the intermediate outputs from the classification network, before and after max-pooling. Afterwards, they fed the aggregated local and global features into the segmentation network. Later, Qi et al. [25] introduced *PointNet++* model, which has similar key structure as the previous PointNet, but improves the model by utilizing a hierarchical structure, similar to the one used in image processing convolutional neural networks. It recursively applies PointNet on a nested partitioning of the input point cloud, starting from small local patches and gradually extending to bigger regions. The whole hierarchical feature learning architecture (on 2D point set example) suggested in the paper is shown in Figure 4.3.

In another study, Wu et al. [34] presented a new convolution operation called

PointConv, which can be applied on unordered and irregular point clouds. They treat convolution kernels as nonlinear weight and density functions of the local coordinates of 3D points. The weight functions are learned with multi-layer perceptron networks and density functions through kernel density estimation. Such learned kernels can be used for translation-invariant and permutation-invariant convolutions on any 3D point set.

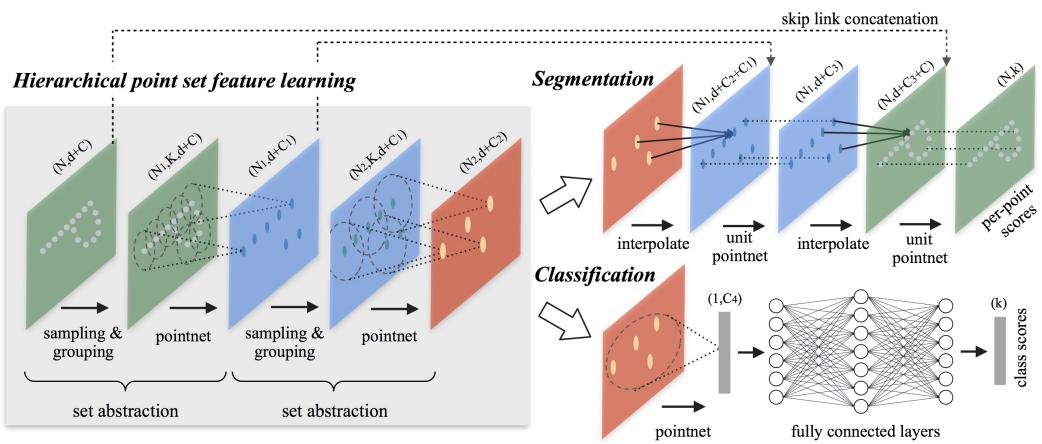


Figure 4.3: The PointNet++ hierarchical feature learning architecture shown on points from 2D Euclidean space [25].

Sparser 3D representations of the human pose, like voxels or point clouds, are usually employed to perform the classification, segmentation, or related tasks. They are rarely used in pose estimation, mainly because the common 2D convolutions cannot be used on this type of data in the same way as on RGB or depth images. Treating point clouds as unorganized sets of points, this type of data can be processed inside the network either by extracting features for each point separately, which yields exclusively local information, or by aggregating the features of all points, which gives us global information about the whole point cloud. Alternatively, the data can be clustered in particular point sets, which are treated as local regions [34]. While in the

classification tasks, the global features are those needed to predict the correct class scores, both local and global information is essential in pose estimation task. Hence, the main issue with performing local context-driven tasks on point clouds is often related to poor propagation of local features inside the network.

Concerning pose estimation task, Ali [1] introduced a novel one-stage approach in his thesis, called *Point-Based Pose Estimation* (PBPE), using point clouds directly as input data to the model which outputs 3D skeleton joint coordinates. He concludes, that since point clouds are able to provide sparser representation of the human body than depth maps, the operations on them would be much easier, and thus, the computational complexity would be reduced. The inspiration for the model was in the PointNet architecture. Besides the proposed PBPE model, the contribution of his work also consists of the refinement of several two-stage methods by using an automatic annotation mechanism for labeling body regions in real data. Next, the study presents the benefits of fusion of the real training data and more complex synthetic training data. The poses in the synthetic dataset are much more varied, so by adding certain amount of the synthetic data to the real dataset during the training phase, they extend the diversity of the training set. As a result, the model is able to generalize better. On the other hand, the synthetic data is also useful for pre-training a model, reducing the computational cost and time of the real data annotation. Thus, such pre-trained model can be fine-tuned on a relatively small part of the real dataset, yet achieving reasonable results.

As a part of our previous research, we re-implemented the method from [1], while slightly modifying the model architecture to improve the final estima-

tions. We enhanced the part of the network which extracts local features of the input point cloud, and reduced the amount of batch normalization in the model.

Chapter 5

Proposed implementation

In this chapter, we are going to describe the implementation we propose, the technologies we intend to use and the models we aim to examine.

5.1 Technologies

In this thesis, we implement all of the studied models in python deep learning library Keras with Tensorflow backend. All of the experiments are conducted on NVIDIA GTX 1070.

5.2 Benchmark datasets

In this section, we are going to describe the datasets which will be used for the evaluation purposes of this thesis. All of the datasets mentioned below were made publicly available for research purposes.

5.2.1 ITOP

The Invariant-Top View Dataset (ITOP) [6] consists of approximately 50K real-world depth images from two camera viewpoints (front view and top view). It captures 20 people, each performing 15 different actions. The dataset comes with the initial partition of the data into the train and test set. The labeled train and test data contain around 18K and 5K samples from each viewpoint, respectively. Besides the depth images and the ground truth joint labels, the dataset also includes raw point clouds. The skeletal model in this dataset is described by 15 body joints. Sample depth maps from the dataset are shown in Figure 5.1. As can be seen in the figure, the depth images are rather noisy, but the noise can be partly reduced by several background segmentation methods.

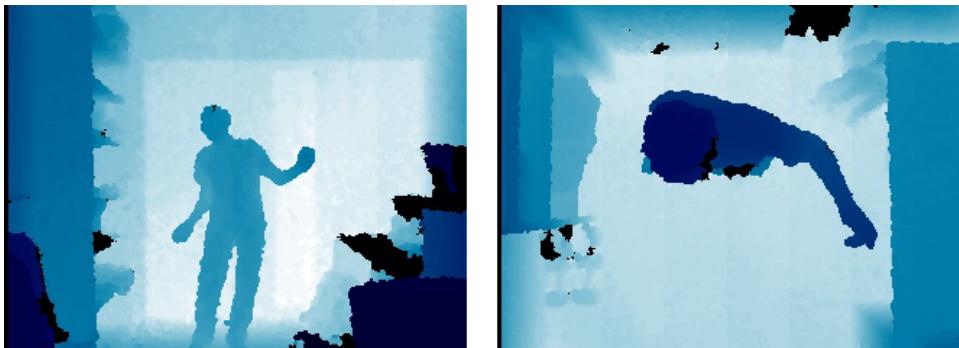


Figure 5.1: Sample depth images from the ITOP dataset [6] (front and top view).

The current state-of-the-art on ITOP dataset was claimed in [14], where the 100% accuracy at 10cm precision was reached, which means every joint was predicted within 10cm from its ground truth position. In the paper, they also claim to have reached the mean error of approximately 0.9cm on ITOP dataset (0.19cm on the front-view data), which is the best result achieved on

the stated dataset, up to our knowledge.

5.2.2 UBC3V

The UBC3V [29] is a synthetic dataset made for the task of pose estimation from multiple cameras. It contains around 6 million synthetic depth frames structured in three parts according to the complexity of the human postures – easy, medium and hard pose. The pose in each frame is represented by the position of 18 skeletal joints. It captures a total of 16 characters and each frame is observed from three different viewpoints. Therefore, the samples can be treated both as multi-view (point clouds from three cameras merged into one), or single-view postures (each sample handled separately). Although the multi-view point clouds describe a particular posture in a more complex way, considering the use-case in most of the real-time applications, the single-view samples are often the preferred option, due to the difficulty of the task of multiple camera synchronization.

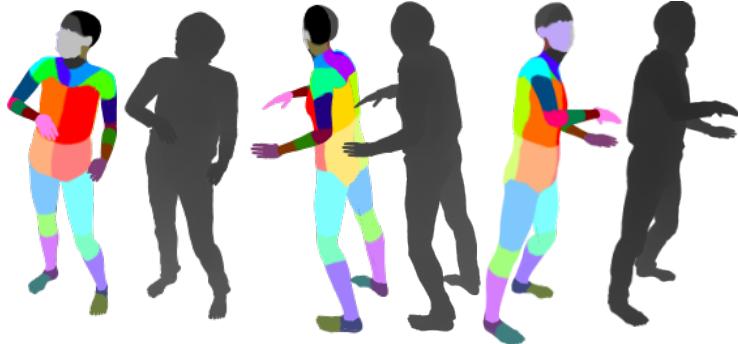


Figure 5.2: Sample data from the UBC3V dataset (the figure shows the same pose from three different cameras) [29].

The depth data are available in the form of depth images, but can be converted into point clouds in world reference coordinates using the intrinsic

and extrinsic camera parameters. The ground truth labels are composed of 18 joints per posture, and the dataset also comprises a segmentation of each point cloud into 43 body regions (as can be seen in Figure 5.2).

5.2.3 MHAD

The Berkeley Multimodal Human Action Database (MHAD) [33] was recorded on real humans. It contains 11 actions performed by 7 male and 5 female subjects. Each subject performed each of the actions 5 times, which yields about 660 action sequences corresponding to about 82 minutes of total recording time. The total number of depth frames is over 250K. The skeleton structure in this dataset is defined by 35 joint locations. Two of the performed actions involve a chair, used for the subject to sit down and stand up. It is worth a mention, that the chair itself provides a lot of clutter in the depth data. The whole dataset was captured by multiple devices, including cameras, depth sensors, accelerometers and microphones. Two kinect cameras were used and synchronized to acquire the depth data, one placed in front of the subject, the other one at the back. Intrinsic and extrinsic camera parameters are provided to extract the scene as a point cloud with real-world coordinates.

5.2.4 CMU Panoptic Dataset

The CMU Panoptic [10] is a massive multi-view dataset containing video recordings from 480 VGA cameras and more than 30 HD cameras, RGB and depth data from 10 Kinect v2 sensors and 3D body poses. The facial and hand keypoint data is available as well. The skeleton representation including the face and hands consists of 19 body joints. The full dataset yields around 6 hours of recordings. The synchronization of the devices is hardware-based, although, as the authors state in the database description, there is no way to

perfectly synchronize multiple Kinects. However, most of the data is aligned accurately by hardware modifications for time-stamping. The calibration parameters are provided for all the cameras, therefore 3D point clouds in real world coordinates can be easily generated. The multi-view point clouds are created by merging single-view data from the multiple Kinects captured within a time interval of ca. 15 milliseconds. If the human motion is fast, there exist certain misalignments, but they can be further filtered according to the delivered synchronization tables, in order to leave out the misaligned frames.

During our study, we utilize only certain parts of the CMU Panoptic dataset, since we are strictly concentrating our research on a single person pose estimation, while the CMU data also contain sequences capturing multiple people in a single frame. Aside from that, for some of the sequences, only certain types of data are included. Specifically, we use the 'range of motion' section of the database, since it is the only section restricted to a single actor at a time, which involves the ground truth 3D poses. Also, we construct the 3D point clouds from only 3 out of 10 Kinect cameras, as the complete set of depth sensors generate extremely dense data, which is not likely obtainable with common setup. Besides, the three Kinects capture the scene within an angular range large enough to yield a full 3D body pose. Since we are not focusing on face keypoints and detailed hand pose estimation, we exclude the corresponding joints from the skeleton structure, what leaves us with a set of 15 body joints.

The database captures multiple actors of different gender, age and body shape. Given the fact the dataset has been recorded in a special environment of the Panoptic studio, the amount of noise in the scene is very limited.

However, the surrounding walls and the floor still need to be removed in the pre-processing stage. The CMU Panoptic dataset does not come with any train and test split, so for the sake of our experiments, we established the test set as random 20% of the data. From the remaining train samples, another 20% are used as a validation set.

5.3 Approaches

The goal of our study is to implement several different approaches to the task of human pose estimation using neural networks. Among the methods we implement are both existing state-of-the-art neural models, as well as our novel approach we propose in this thesis – a two-stage pipeline called Segmentation-Guided Pose Estimation. We present the existing models we re-implement, and our novel approach in the subsequent sections.

5.3.1 Existing models

Deep Depth Pose model

One of the aims of this work, is to re-implement the Deep Depth Pose model (DDP) [14] in Keras framework. We consider this an essential step to propose a state-of-the-art pose estimation model performing on depth data, since the DDP model claims to outperform all of the present methods on the examined datasets. Originally, the model was implemented in Matlab and while some parts of the code (mostly testing) are publicly available, the training procedures have not been published.

The main idea behind the stated method is incorporating a set of predefined prototype human poses. The convolutional neural network outputs a vector

containing weights, each corresponding to particular prototype pose. The weighted prototypes are then summed up producing the final estimated pose. The predefined prototype poses should ideally cover the largest possible range of human poses, so that by their linear combination we can obtain any desired pose during the testing. Hence, the set of prototype poses is formed by applying K-means clustering on the training dataset.

Point-Based Pose Estimation model

The Point-Based Pose Estimation model (PBPE) [1] is an approach to 3D human pose estimation from depth data, remarkable by several of its attributes. The whole network consists of two branches, the regression branch and the auxiliary segmentation branch. The output of the model is a 3D body pose, given by the coordinates of skeletal joints. One of the benefits of the method is the possibility of omitting the auxiliary branch in the testing phase, where the model is already trained, and by this restriction of the model's depth and complexity, the computational cost and prediction time can be cut down.

Of great significance for our thesis is the fact, that the input to the PBPE network is directly in a form of an unstructured point cloud. This may be convenient for a number of reasons, one of them being the acquisition of a sparser representation of the human body pose, in comparison to commonly used depth maps. The more density the input covers, the more computations have to be performed in the network, therefore point clouds offer more effective data processing even in complex model architectures. Also, by working directly with point clouds, we omit the perspective distortion present in depth maps, and therefore avoid the need for the network to perform distortion-invariant estimation. The model architecture is inspired by

PointNet model [24]. Inside the PointNet as well as the PBPE model, the point clouds are processed using the pseudo-convolutions with kernel size 1×1 , which are generally employed to change the filter space dimensionality, mostly reducing the number of depth channels.

While being originally implemented in pure Tensorflow, the source code for the PBPE model has not been published at the time of our research. We aim to implement the stated model in Keras framework from scratch and evaluate the results on benchmark datasets.

5.3.2 Segmentation-Guided Pose Estimation

An important part of our work is to develop a novel method for human pose estimation. We introduce the Segmentation-Guided Pose Estimation (SGPE) – a two-stage pipeline which takes a point cloud as an input, and outputs the 3D coordinates of the estimated skeletal joint positions. Incorporating the idea of handling unorganized and permutation-invariant point clouds, both stages of the pipeline are based on pseudo-convolutions, which operate in the filter dimension. The first stage of our pipeline involves a segmentation network, which classifies the points representing a human pose into the corresponding body regions. In the second stage, the original input point cloud containing the point coordinates is concatenated with the output regions from the segmentation network, thus forming a four-channel point cloud input. Such produced data, conserving together the local as well as the global information, is then fed into the second model – the regression network, where the joint coordinates are finally regressed. The architecture of both networks make use of residual connections added to the shared multi-layer perceptron blocks, to strengthen the feature propagation.

Similarly to the models mentioned above, we implement our novel approach in Keras framework as well. To be able to compare the performance of our model to that of the existing ones, the same benchmark datasets are used for the evaluation.

Chapter 6

Implementation

This chapter is dedicated to the description of the re-implementation of several existing models, along with our modifications, and the implementation details of the novel proposed method. We will introduce the architecture of the models, as well as the data pre-processing, training procedures and optimization parameters used.

6.1 Re-implementation of existing models

6.1.1 Deep Depth Pose model

In this section, we are going to describe our re-implementation of the DDP model (proposed in [14]) in Keras framework. The follow-up additional modifications of the model, which are presented at the end of the section.

The DDP model is based on the idea of linearly combining the predefined prototype poses to obtain the resulting pose estimation. The set of prototype poses is produced by clustering the training dataset into K clusters, K being a hyperparameter of the network. The output of the model repre-

sents K weights, each corresponding to one of the prototype poses. The final estimated pose is then obtained as a sum of the weighted prototypes.

Model structure

The architecture of the DDP model itself is relatively simple. The model is actually not very deep, it consists of five convolutional blocks, each followed by ReLU activation. First three blocks also contain a pooling layer, as indicated in Figure 6.1. After the convolutional blocks, there are three fully-connected layers. The number of neurons in the output layer is a hyperparameter set to the number of prototype poses clustered from the dataset. The input to the model is a one-channel depth map of size 100 x 100 pixels.

<i>Input</i>	Conv01	Conv02	Conv03	Conv04	Conv05	Full01	Full02	Full03
100 x 100 P: 2 x 2	7 x 7 x 96 P: 2 x 2	5 x 5 x 192 P: 2 x 2	3 x 3 x 512	2 x 2 x 1024	2 x 2 x 2048 Dr=0.2	1024	256	K

Figure 6.1: Deep Depth Pose model architecture [14] (P stands for pooling layer, Dr indicates dropout, K is the number of prototype poses).

Loss function

The loss function used during training the model is defined as huber loss with a regularization term, as described in Equation 6.1, where \mathcal{L}_R is the (residual) huber loss.

$$\mathcal{L}_{DDP}(g(\mathbf{D}, \theta), \mathbf{C}, \mathbf{p}, \alpha) = (1 - \alpha) \cdot \mathcal{L}_R(\mathbf{C} \times g(\mathbf{D}, \theta), \mathbf{p}) + \alpha \cdot \|g(\mathbf{D}, \theta)\|_1 \quad (6.1)$$

In the equation above, $g(\mathbf{D}, \theta)$ represents a non-linear function computed

by the neural model on the input depth map \mathbf{D} with trainable parameters θ , returning a column vector of length K (number of clustered prototype poses), \mathbf{C} is a matrix with K columns containing prototype poses, \mathbf{p} stands for the vectorized ground truth pose, and α is a regularization coefficient – a hyperparameter to control the magnitude of the resulting weights of the prototypes.

Initialization

The values of the model’s hyperparameters from the original DDP paper were chosen using the ‘Mann-Whitney U-test’ [13], thus determining, whether the median difference between pairs of configurations are statistically significant. The value of the regularization coefficient was fixed to $\alpha = 0.08$ for the ITOP dataset, and $\alpha = 0.01$ for the UBC3V dataset. The number of prototypes was set to $K = 70$ and $K = 100$ regarding the ITOP and the UBC3V dataset respectively.

The model learns using the Adam optimizer with the learning rate set to 10^{-3} , which is progressively decreasing during the training. The weights in the model are initialized using the Xavier normal initializer, that is randomly chosen from the normal distribution with zero mean and a standard deviation proportional to the filter size. The biases are initially set to zero. The size of batches is set to $b = 64$.

Data pre-processing

In the pre-processing stage, the depth maps are re-sized to match the model input dimensions. First, the image is cropped along the larger dimension on both sides to fit the square, assuming the subject is located near the center

of the image. Then, the image is re-sized to 100×100 px. Afterwards, the depth values are normalized to range $[0,1]$. Also, the input depth images are normalized globally by subtracting the mean image of the training set from each depth sample. The ground truth poses are being normalized as well – to zero mean and one standard deviation.

Concerning the datasets used for evaluation, UBC3V already comes with the split into the train, validation and test set. The ITOP dataset is originally divided into the train and test set only, thus the validation set was acquired by random sampling of the train data. The ratio was set to 80/20, meaning the data used for validation are representing 20% of the train set. During the experiments, only the hard-pose part of UBC3V, and the side-view part (for single-view approach) of the ITOP dataset were used.

Modifications

This section introduces several modifications we have made to the proposed DDP model, in order to lower the mean error of the predictions. Since the experience of the researchers made us believe that the implementation of the same model in different framework often leads to different results, we diverted from the original architecture of the model, as well as the initial optimization parameters.

However, still inspired by the DDP model, we have built a slightly deeper, more complex model, in an attempt to improve the final estimations, assuming the original architecture was not able to fully retain the complexity of the specified task. We used both ITOP and UBC3V datasets again, for the evaluation of the modified model. The architecture of the modified DDP network is shown in Figure 6.2. As shown in the figure, we significantly increased the

amount of dropout layers, as the model tended to overfit the training data. Moreover, two additional convolutional layers were stacked in the model architecture, to enhance the nonlinearity of the network (introduced by the ReLU activations used after convolutions).

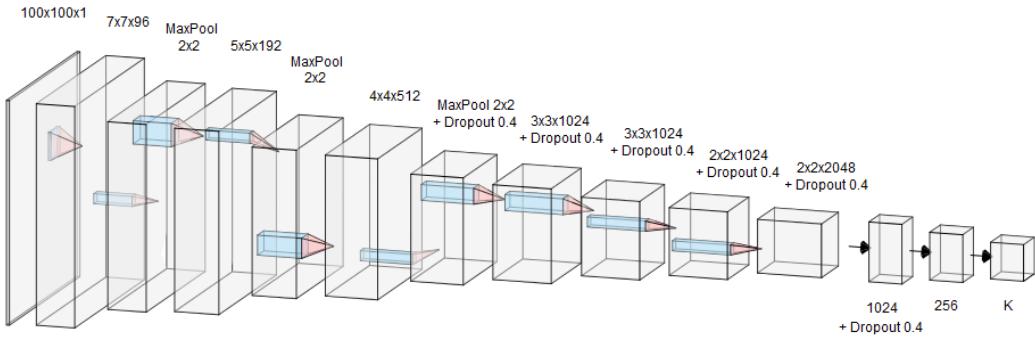


Figure 6.2: Our modification of the original DDP model architecture (K stands for the number of prototype poses).

Besides the structure of the model, some of the hyperparameters were slightly modified as well, mainly due to switching to another deep learning framework, where the back-end procedures differ, which may be resulting in a convergence to different outputs. Based on the experiments, the regularization parameter was fixed to $\alpha = 0.1$ for both benchmark datasets. The residual huber loss has been replaced by an absolute error, and the L2 norm was used in the regularization term instead of the L1 norm. Even though in the original paper, the authors indicate the L1 norm is more robust to outliers, our tests in the utilized framework led to a conclusion, that using the L2 norm as a regularization term in the loss function was more helpful to the loss minimization. Also, the initial learning rate was reduced to 10^{-4} , due to an occasional undesirable divergent behavior of the model loss.

Regarding the test phase, the experiments on the ITOP dataset have shown that even though the test set seems sufficiently representative (that is, retains the approximate distribution of the training set), the error on the test data is by far worse than the error on previously unseen data obtained by random sampling of the training set (used as a validation set). This is probably caused by the fact, that the input data is formed by a sequence of video frames, thus many subsequent samples captures almost the same content, with only a minimal shift between the frames. As a result, a lot of the validation samples are nearly the same as the ones used for training.

The final results of both the original DDP and modified DDP model, evaluated on the test set of ITOP and UBC3V datasets, are reported and discussed in Chapter 7.

6.1.2 Point-Based Pose Estimation model

As another part of our study, we chose to examine the Point-Base Pose Estimation (PBPE) model introduced in [1]. Again, we re-implement the method in Keras framework, and propose further modifications of the model at the end of the section.

The model is heavily inspired by the implementation of PointNet [24], which was built for the task of object classification and semantic segmentation. Basically, the architecture has been modified to fit the task of human pose estimation, making use of the auxiliary sub-network contributing to the global model loss.

Model structure

The architecture of the proposed model consists of two branches (or sub-networks). The auxiliary sub-network is included to compute the body part segmentation on the fly, while the main network regresses the joint locations. The basic idea behind the structure of the model is the aggregation of both local and global features of the input point cloud in the auxiliary part-segmentation network. Without the sub-network, almost all of the local context would be lost because of the max-pooling operation in the intermediate layers. The incorporation of the local features helps the network understand the relationships among particular local regions of the human body. The whole architecture of the stated model is shown in Figure 6.3.

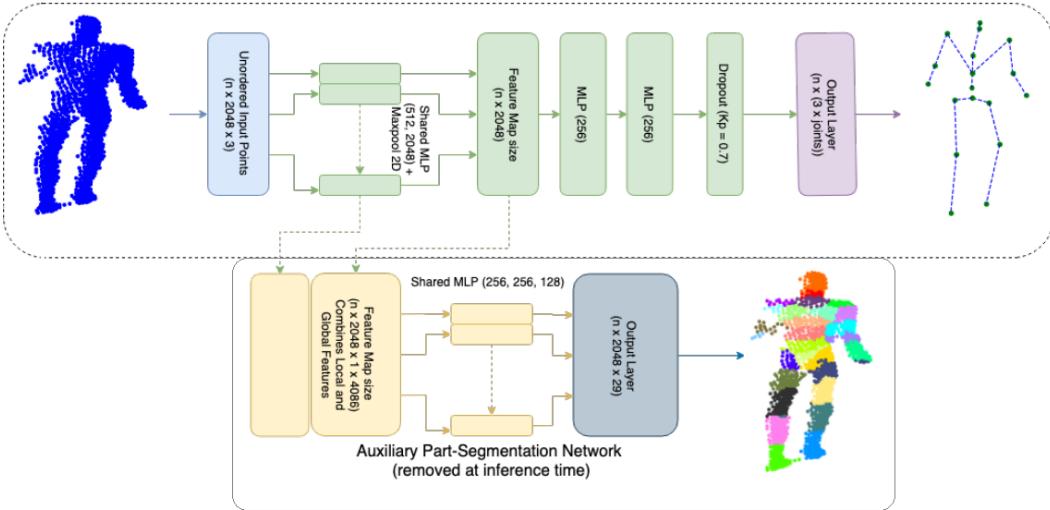


Figure 6.3: The Point-Based Pose Estimation model architecture [1].

As already mentioned, the unorganized and irregular depth input represented by a point cloud is a key factor in this architecture. For this very reason, the engagement of the classic convolution operation is not possible, since

the point cloud provides no explicit spatial information about the neighboring points. In this case, pseudo-convolutions with a kernel size of 1×1 are employed. The functionality of such pseudo-convolution layers lies in reducing (or expanding) the dimensionality in the filter space. Shared multi-layer perceptron (shared MLP) stands for a stack of the 1×1 convolution layers, each followed by batch normalization and ReLU activation, as illustrated in Figure 6.3. The numbers in the parentheses indicate the number of filters in particular layers.

After the shared MLP, the regression branch consists of two fully-connected layers and an output layer. Each dense layer is also followed by a batch normalization layer and ReLU activation. The input of the model is of shape $(b \times p \times 3)$, where b represents the batch size, and p is the number of points in the input point cloud, each located by three coordinates. The output shape of both sub-networks depends on the number of joints in the skeleton structure. The number of joints, as well as the number of points the point clouds are containing, is a hyper-parameter.

The input passed to the auxiliary segmentation sub-network is formed by concatenating the local and the global features obtained in the regression branch. The local features contains the outputs of pseudo-convolutions before the max-pooling is applied. The global features are acquired as the result of max-pooling across all points in the point cloud. The auxiliary sub-network outputs per-point labels segmenting the input point cloud into particular body parts.

One of the conveniences of the PBPE model is the optional omission of the auxiliary branch at test time. Since we are focusing strictly on the task

of pose estimation, the body part segmentation is only relevant for us in a context of training (and contributing to the model loss). Hence, we can prevent the data from being passed through the segmentation branch at the inference time, saving computational cost and time.

Loss function

As a loss function of the regression branch, the model uses a simple mean squared error. Regarding the segmentation branch, the categorical cross-entropy is being used, since it is essentially a classification task.

According to our experiments, without the segmentation sub-network, the model tends to slightly overfit the train data. The loss of the auxiliary segmentation branch is contributing to the global loss to help the model generalize better on previously unseen data. Originally, the weight of the contribution was set to $w_a = 0.1$. The default weight of the global loss is $w_g = 0.9$.

Initialization

The weights inside the model are sampled from the normal distribution with zero mean and standard deviation relative to filter size, using the Xavier initializer. Following the PBPE proposal, the initial learning rate is set to 10^{-3} , and is reduced exponentially with a decay rate $d = 0.5$ in each epoch. Also, the learning rate is clipped at the value of 10^{-5} , in order to prevent the model training from getting stuck in local minima. The model is trained using the Adam optimizer. In each step, the model processes a batch of size $b = 32$.

Another essential hyper-parameter is the predefined number of points in the

input point clouds. The value is set to $p = 2048$, which seems to be an adequate amount of points to cover the whole human pose with enough complexity. Therefore, every point cloud needs to be sub-sampled to this number of points, prior to feeding it to the network. The number of joints in the skeleton representation, defining the output shape of the both model branches, should be adjusted according to the dataset the network is processing at the moment.

Data pre-processing

Depending on the single or multi-view approach, the input point clouds are treated each as a single sample, or the different viewpoints of the same frame are merged into a single point cloud. Based on the existing studies, inference from multi-view input data is expected to give more accurate results, however the single-view based estimation is far more useful in most of the real-time applications, since there is no need for camera synchronization. Therefore, we have carried out experiments inferring from multi-view, as well as single-view input data.

As the next pre-processing step, the point clouds were sub-sampled to the specified number of points ($p = 2048$) using the farthest point sampling technique. The input point clouds and the corresponding joint locations were normalized to the range $[-1, 1]$ using the minimum and maximum values of the whole training set. Concerning the validation and test data, there were two options, how to normalize them – either using the scaling parameters of the training set, or scale the validation and test set to the specified range with its own parameters. However, since the model is designed to work in a scenario, where the test data are fed into the network one-by-one, the only

possible option is scaling it using the known parameters of the train set.

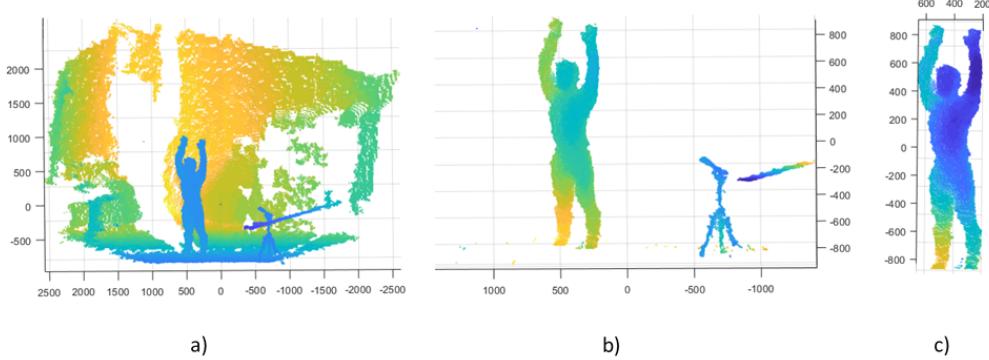


Figure 6.4: Data pre-processing shown on a sample point cloud from MHAD dataset a) before segmentation, b) after removing the background and floor, c) after final clustering.

Since the used datasets (except for the UBC3V) captures real-world data, the raw point clouds contain a lot of noise. Therefore, we consider suitable preprocessing as a very important step. First, the background wall and the floor are removed by MSAC plane fitting (which is a variation of RANSAC algorithm). Then, the segmentation of the point cloud into clusters based on Euclidean distance is performed, and the biggest cluster is considered as the desired subject. Figure 6.4 depicts a sample from MHAD dataset, before and after applying the segmentation pipeline.

The MHAD dataset does not provide any partition into the train and test data. Thus, we carried out experiments using two different protocols: (a) choosing the test set as randomly sampled 25% of the dataset, (b) leave-one-subject-out cross validation. As already pointed out in the original proposal of the PBPE model, the skeleton representation in the MHAD dataset is very detailed, yielding a total of 35 skeletal joints. We have slightly modified the

original skeleton structure by removing several redundant joints – one pair repeated at fingertips, two additional pairs present at toe tips. This way, we restricted the skeleton to the resulting 29 joints (as shown in Figure 6.5), in the same way as in [1]. However, we present results of our approach also on the original full skeleton, to be able to compare our strategy to the existing methods. Since in the case of the modified skeleton we have only removed the redundant skeletal nodes, we have not reduced the complexity of the skeleton in a significant manner, but rather increased the focus on more relevant joints in the skeleton.

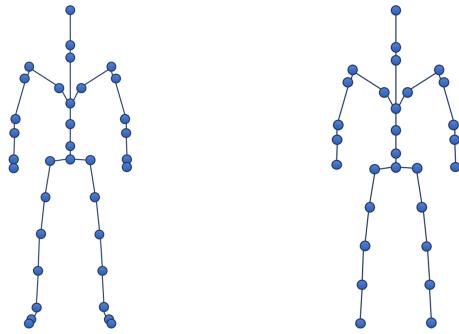


Figure 6.5: The original skeleton structure used in MHAD dataset (*left*) vs. the modified skeleton (*right*).

As the real datasets do not comprise the partition into the particular body regions, the automated technique for the annotation has been proposed in the original implementation of PBPE model. Each skeletal joint is assigned one body region. Using the automatic annotation, each point of the point cloud is associated with the closest skeletal joint based on Euclidean distance. The example of the annotated point cloud is shown in Figure 6.6.

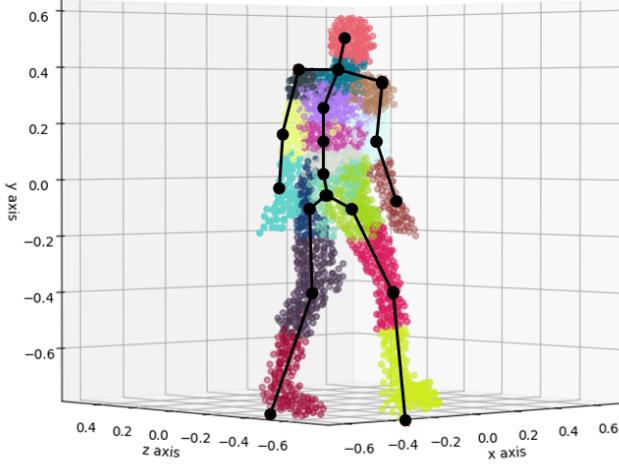


Figure 6.6: The automatic body region annotation on a sample real data point cloud.

Modifications

For the sake of achieving as accurate pose estimation as possible, we made a number of modifications to the model implementation, to use the full potential of the selected framework. In the architecture of our modified PBPE model, we put even greater emphasis on the local feature extraction, chaining additional pseudo-convolutional layers into the shared MLP before max-pooling. Aside from that, we also left out the dropout in the regression branch, since with a help from the auxiliary branch, the model had not suffer from overfitting. Moreover, we skipped all the batch normalization layers in the main branch, since the performance improved when excluding them.

The loss function has been replaced by mean absolute error, as it helped the model to converge faster in our case. The hyperparameters of the model mostly preserve the same values as in the original paper, except for the

learning rate and loss weights. Given by our observations, we fine-tuned the weight parameter of the global loss to $w_g = 1$, and decreased the weight of the segmentation branch to $w_a = 0.01$. We set the initial learning rate to 5^{-4} , while reducing it with an exponential decay rate of $d = 0.2$ in each epoch. Again, the differences between the performance with the particular hyper-parameter values might be caused by switching to another deep learning framework.

6.2 Segmentation-Guided Pose Estimation

This section is dedicated to the implementation of a novel approach we present as one of the contributions of this thesis. The core of the proposed method is a two-stage pipeline, consisting of a body segmentation network and a regression network. The two subsequent stages take a point cloud as an input and produce skeletal joint coordinates as a result.

The aim of the model is to embrace both local and global features of the input point cloud, and thus increase the accuracy of the final estimation of the human pose. Also, the model makes use of the residual connections in-between the layers to help the gradient flow, and to avoid the degradation problem as the network depth increases.

Model structure

In both networks of the pipeline, the input point clouds are being processed in the similar fashion as in the PBPE model, i.e. by employing the shared multi-layer perceptron modules. The architecture of the whole pipeline is described in Figure 6.7.

As a first part of our pipeline, we propose a segmentation network with an architecture similar to the one of the regression model, instead of making use of one of the existing segmentation methods. While utilizing the same main modules in the segmentation and regression model, we believe preserving a similar network-specific representation of the body pose in both models works for the benefit of more accurate pose estimation.

In the first stage, the pre-processed point clouds are fed into the segmentation network, which performs a pointwise classification into the corresponding body regions. The architecture of the model, as shown in Figure 6.7 (top), is based on the shared MLP modules. To obtain global features, the output vector of the first shared MLP is aggregated in a pooling layer across all points of the point cloud. Since the local information is essential in the task of semantic segmentation as well, we want to avoid losing the local context after the max pooling aggregation. Therefore, the local features extracted from the intermediate layers of the shared MLP are concatenated with the aggregated global features and sent off to the second shared MLP module. After the second shared MLP, the model outputs the predicted per-point classification probabilities for each body region.

In order to help the gradient flow, and enhance the feature propagation, we improved the shared MLP modules in our approach by adding residual connections in-between the convolutional layers. Referring to the figure, the numbers in the brackets near the shared MLP blocks describe the number of filters in the respective 1×1 convolutional layers.

Since the real data does not come with body-parts segmentation, we perform an automatic annotation of the point clouds to acquire ground truth body

region classification of the data, in the same way as described in PBPE model implementation.

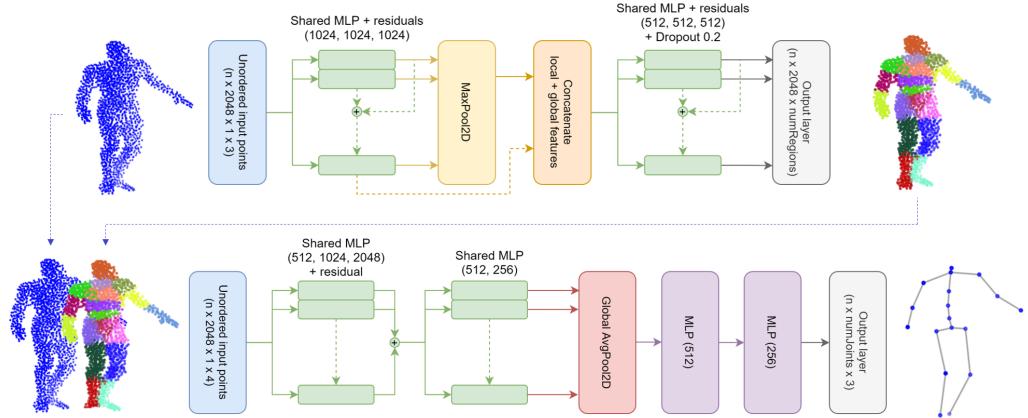


Figure 6.7: The Segmentation-Guided Pose Estimation pipeline: First, point clouds are segmented into body regions in the segmentation network (*top*), then the input point clouds are concatenated with the predicted per-point body region assignment as a fourth channel, and fed into the regression network (*bottom*).

The second stage of our pipeline is based on the regression network. To incorporate the idea of retaining both local and global context of the input point clouds, the initial 3D point cloud is concatenated with the predicted pointwise region assignment after the body region segmentation, forming a four-channel input point cloud, which is passed to the regression model (as indicated in Figure 6.7, bottom). Again, the network incorporates two shared MLP blocks. The first one contains three convolutional layers with 1×1 kernels, followed by one residual connection adding up the outputs of the three preceding layers. To control the number of parameters of the network, the second shared MLP includes two layers and no additional skip connections. To avoid having majority of the model parameters concentrated in the first fully-connected layer, the global average pooling is utilized instead

of a simple flattening layer to spatially average across all points right before the fully-connected layers. As indicated in the figure, the pooling layer is followed by two dense layers with 512 and 256 filters, respectively. Finally, the model estimates the 3D skeletal joint coordinates of the captured human subject as the output.

Loss function

In the first (segmentation) network, the categorical cross-entropy is employed as a loss function, to measure the accuracy of the body part classification. In the case of the regression network, the mean absolute error between the predicted locations and the ground truth labels of all skeletal joints is used to determine the model loss.

Initialization

Both networks are trained using the Adam optimizer with the initial learning rate equal to 10^{-3} , and an exponential decay rate of $d = 0.2$ applied at the end of each epoch. All weights are initialized with Xavier normal initializer. The batch size is fixed to $b = 32$ for both models.

Data pre-processing

Prior to sending the input point cloud to the first neural network, the background scene is segmented out – the ground floor and the surrounding walls are removed using RANSAC plane fitting algorithm, and the biggest cluster of the point cloud is extracted, being considered the captured human subject. To unify the dimension of the model input, the point cloud is subsampled to a fixed number of points using the farthest point sampling. We set the hyperparameter determining the number of points in each point cloud to

$p = 2048$, yielding a fair density of the input data. Both the ground truth skeleton coordinates, as well as the input point clouds, are normalized to the range $[-1, 1]$ along each axis, using minimum and maximum values of the whole training set.

Chapter 7

Results

In this chapter, we present results of our proposed approach evaluated on benchmark datasets. We will go through the evaluation metrics and particular experiments conducted using various evaluation protocols. Moreover, we will compare our results to existing state-of-the-art methods.

7.1 Evaluation metrics

In the process of evaluation, we used mean per joint position error (MPJPE) and mean average precision (mAP) as metrics, following [1, 6, 14, 29]. Mean average precision is defined as percentage of all skeletal joints predicted under 10 cm threshold from their ground truth position.

7.2 Experiments

For the purpose of evaluation, we used several benchmark datasets, including the challenging ITOP front-view [6], UBC3V hard-pose [29], MHAD [33] and a subset of CMU Panoptic dataset [10]. On a test set of the ITOP front-view

dataset, the mean per joint position error our method achieves is 6.40 cm (as shown in Figure 7.1, left). Using a 10 cm threshold, the mean average precision is 85.57%, which is comparable to the state-of-the-art results.

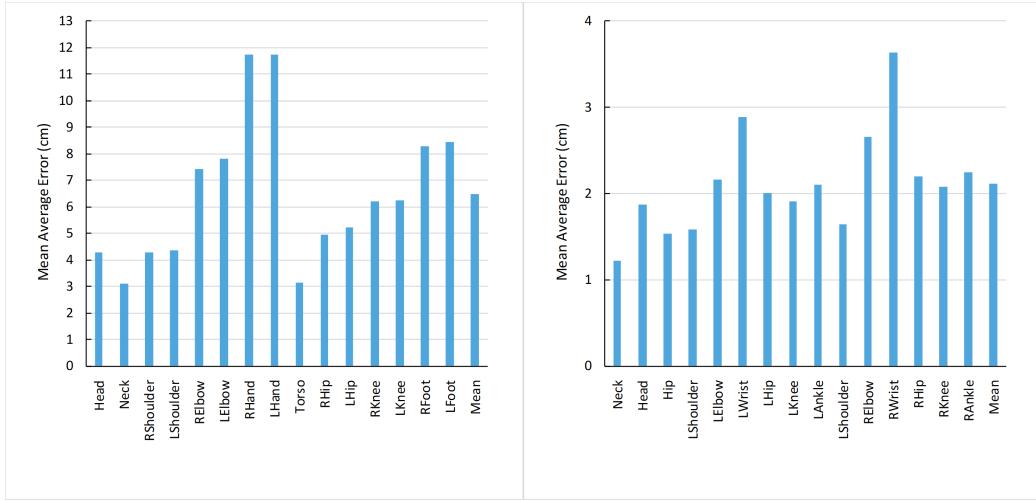


Figure 7.1: Mean average error per joint on ITOP (*left*) and CMU (*right*) dataset.

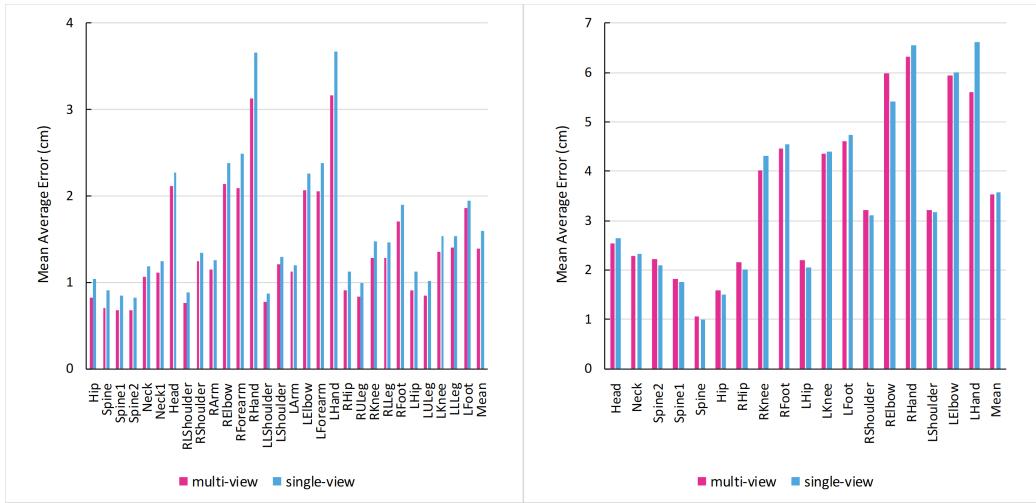


Figure 7.2: Mean average error per joint on MHAD (*left*) and UBC3V (*right*) dataset, comparing multi-view and single-view approach.

Regarding the CMU dataset, we evaluated our method specifically on the *Range of motion* section of the dataset, yielding approximately 141K frames, as it was the only section capturing a single person, having ground truth labels available at the time of this research. Since prior to our work, there was no protocol established for the utilized section of the dataset, and considering the amount of data in the selected section of the dataset, we marked 20% of the data obtained by random sampling as the test set. There are also no existing results to compare to, concerning the single person pose estimation on this dataset (up to our knowledge). The mean per joint position error using our proposed approach is 2.11 cm (as shown in Figure 7.1, right), and the mean average precision at 10 cm is 98.39%. Figure 7.3 illustrates the qualitative results on samples from CMU Panoptic dataset.

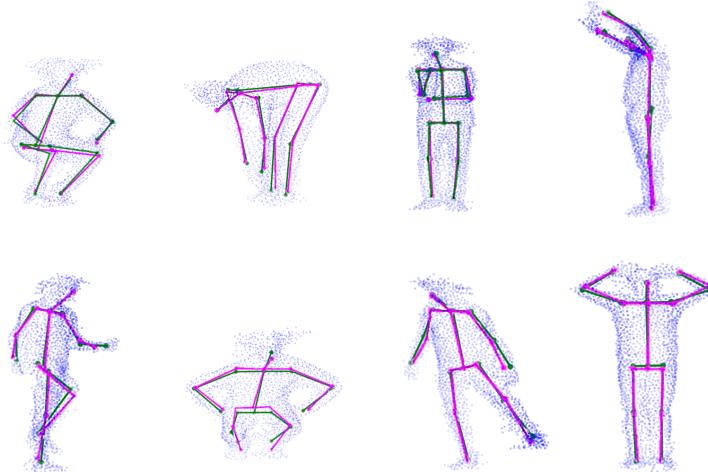


Figure 7.3: Qualitative results of our method on CMU Panoptic dataset. The ground truth skeletons (*green*) vs. our estimation (*magenta*). Best viewed in color.

Similarly, the MHAD dataset does not originally come with a train and test split, thus we carried out experiments using two different protocols: (a)

choosing the test set as randomly sampled 25% of the dataset, (b) leave-one-subject-out cross validation. We present results of our approach on both the original full skeleton (containing 35 joints) and the modified pruned skeleton (29 joints) to be able to compare our strategy to the existing methods (as shown in Table 7.1). Since in the case of the modified skeleton we have only removed redundant skeletal nodes, we have not reduced the complexity of the skeleton in a significant manner, but rather increased the focus on more relevant joints in the skeleton. As can be seen in Table 7.1, the mean per joint position error has visibly decreased after omitting the redundant skeletal joints.

Table 7.1: The mean per joint position error (MPJPE) of our approach on MHAD dataset evaluated following the leave-one-subject-out (LOSO) cross validation strategy, as well as randomly sampled test set, compared to state-of-the-art methods.

Method	eval. protocol	MPJPE (cm)	MPJPE (cm)
		single-view	multi-view
Shafei et. al [29]	LOSO	–	5.01
PBPE [1]	random 25%	7.46	3.92
PBPE [1] (29 joints)	random 25%	3.20	–
Ours - FCPE	LOSO	3.97	3.36
Ours - FCPE (29 joints)	LOSO	3.23	2.97
Ours - FCPE	random 25%	1.85	1.62
Ours - FCPE (29 joints)	random 25%	1.59	1.39

It is important to point out that after the visual inspection of the MHAD dataset, the ground truth position labels in certain frames or sequences are noticeably shifted from the location of the corresponding skeletal joints, and clearly erroneous. Consequently, the model might overfit the training data, as in some cases the estimated joint position is visibly closer to the real joint

location than the ground truth label.

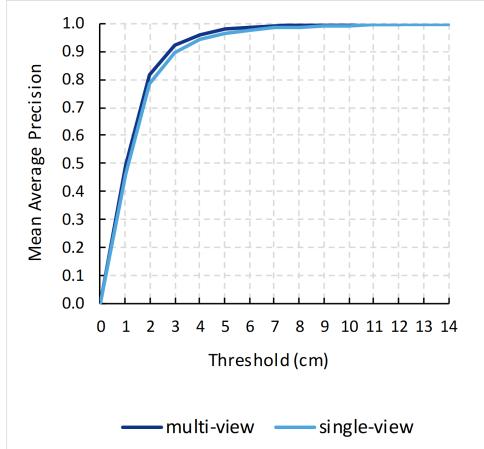


Figure 7.4: Mean average precision at threshold on MHAD dataset for multi-view and single-view approaches.

Following the first protocol, i.e. establishing the test set as 25% of the data by random sampling, our method achieves the mean per joint position error as low as 1.39 cm for the multi-view approach, and 1.59 cm for the single-view approach (as shown in Figure 7.2, left), when using the modified skeleton structure. The achieved mean average precision at 10 cm is as high as 99.80 % and 99.21 % for the multi-view and single-view approach, respectively (Figure 7.4). We set a novel state-of-the-art for MHAD dataset, lowering the mean per joint position error by almost 65% following the multi-view approach, and by approximately 50% following the single-view approach.

Table 7.2 summarizes the mean per joint position error on UBC3V hard-pose dataset for both single-view and multi-view approach. Using our approach, the achieved mean per joint position error is 3.57 cm in the case of single-view data, and 3.53 cm using multi-view data (as shown in Figure 7.2, right). The mean average precision at 10 cm is 95.63 % and 95.71 % for the single-

view and multi-view approach respectively. The claimed results of the Deep Depth Pose (DDP) model proposed in [14] are listed in italics, due to a number of unsuccessful attempts to reproduce them by various researchers. The observed results on the reproduced DDP model, implemented following the same training procedures as the original implementation, are indicated in the table as well. Sample qualitative results on UBC3V hard-pose test set are shown in Figure 7.5, predicted on merged multi-view point clouds.

Table 7.2: The mean per joint position error (MPJPE) of the proposed method on the test set of the UBC3V hard-pose dataset compared to state-of-the-art methods.

Method	MPJPE (cm) single-view	MPJPE (cm) multi-view
DDP (observed)	19.23	–
PBPE [1]	7.59	5.59
Shafei et. al [29]	–	5.64
DDP (claimed) [14]	<i>3.15</i>	<i>2.36</i>
Ours - FCPE	3.57	3.53

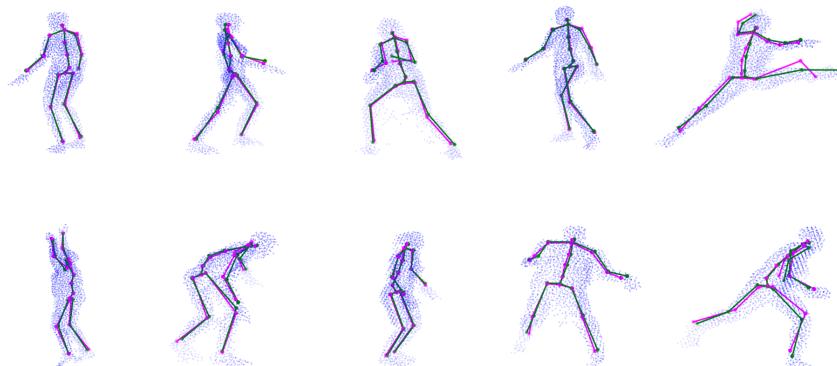


Figure 7.5: Qualitative results of our approach on test set of UBC hard-pose dataset. The ground truth skeletons (*green*) vs. our estimation (*magenta*). Best viewed in color.

We also present evaluation of the first stage of our pipeline. The accuracy of the semantic segmentation into the corresponding body regions over training epochs for all examined datasets is depicted in Figure 7.6. Our method achieves up to 95% segmentation accuracy on CMU Panoptic dataset.

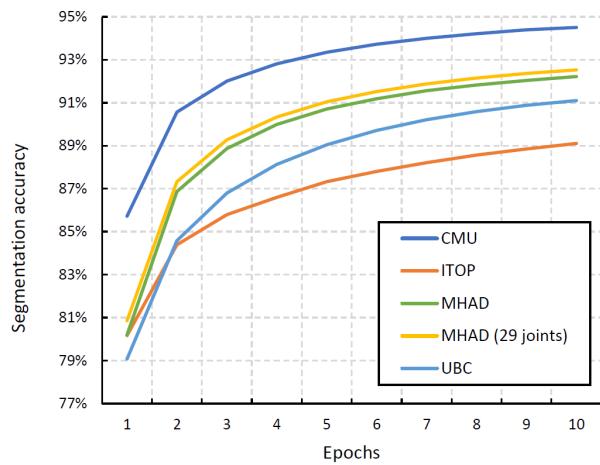


Figure 7.6: Accuracy of the per-point body-parts segmentation performed in the first stage of our pipeline on all examined datasets.

Chapter 8

Conclusion

In this thesis, the task of 3D human pose estimation from depth data is studied, while the core of the research is focused on the impact of a particular form of input data on the estimation accuracy and computational efficiency.

As the main contribution of our work, we propose a novel two-stage deep learning method for an accurate single-person depth-based human pose estimation called Segmentation-Guided Pose Estimation. We eliminate drawbacks related to the projection of 3D space to a 2D image, when estimating pose from depth maps, by introducing a concept of unordered point clouds as a permutation-invariant input to a neural network. To allow the network to maintain both local and global contextual information, we employ intermediate concatenation of extracted pointwise and aggregated features inside the model. Additionally, we perform semantic segmentation of the input point cloud into the corresponding body regions, and utilize the per-point region assignment as an extend of the input point cloud before the final regression. Our proposed approach proved to benefit from the redundancy of the input to the segmentation and regression network, being able to absorb the local

and global context at once.

We believe engaging sparse point clouds as an input to the neural network instead of the commonly used depth maps allows us to provide a representation of the human body that is easier to be perceived by the network, while lowering memory requirements and computational cost at the same time. Moreover, to help preserve gradient flow throughout the entire depth of the network, we improved the shared multi-layer perceptron modules by additional skip-connections. Our strategy achieves competitive results on a number of benchmark datasets, and outperforms state-of-the-art approaches.

Our interest to make use of the raw point clouds on the input, and focus on processing them directly, was aroused by the intention to apply the method proposed in this study on data obtained by the 3D MotionCam by Photo-neo [23], which is currently the highest resolution and highest accuracy 3D camera in the world. The 3D camera produces a raw point cloud of the captured scene per frame. Thus, after being subsampled and normalized, the point cloud can be passed directly to our two-stage pose estimation pipeline.

We consider an important part of this study to point out the most relevant limitations we encountered during the experiments. Regarding the depth-based human pose estimation, we see the biggest shortage in the range and accuracy of the available datasets. The suitable public datasets, containing both depth data of a captured human subject and the ground truth skeletal joint coordinates, are either too small to be used as training data for a neural network, or the accuracy of the ground truth labels is not sufficient. Moreover, even in large datasets, the data is often incomplete for certain sections, so the valid subset of the dataset ends up of a too small range after

all. The limited accuracy of the ground truth poses is usually caused by poor synchronization of a depth sensor and a motion capture system. The most commonly used depth sensors do not have a stable frame rate, which results in time delays and misalignment between frames, and makes the precise synchronization practically impossible. In some of the datasets, this issue is partly fixed by time-stamping technique, refining the frame alignment, and filtering out the mismatches. It is even harder considering the multi-view approach, when the multiple depth sensors need to be synchronized mutually as well as with the motion capture system.

Bibliography

- [1] Ali, Ashar. 2019 (May). *3D Human Pose Estimation*. MS Thesis, Georgia Institute of Technology.
- [2] Chen, Ching-Hang, & Ramanan, Deva. 2016. 3D Human Pose Estimation = 2D Pose Estimation + Matching. *CoRR*, **abs/1612.06524**.
- [3] Chou, Chia-Jung, Chien, Jui-Ting, & Chen, Hwann-Tzong. 2017. Self Adversarial Training for Human Pose Estimation. *CoRR*, **abs/1707.02439**.
- [4] Ge, L., Liang, H., Yuan, J., & Thalmann, D. 2017 (July). 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images. *Pages 5679–5688 of: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [5] Ge, Liu-hao, Liang, Hui, Yuan, Jun-song, & Thalmann, Daniel. 2016. Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs. *CoRR*, **abs/1606.07253**.
- [6] Haque, Albert, Peng, Boya, Luo, Zelun, Alahi, Alexandre, Yeung, Serena, & Fei-Fei, Li. 2016 (October). Towards Viewpoint Invariant 3D Human Pose Estimation. *In: European Conference on Computer Vision (ECCV)*.
- [7] Huang, Fuyang, Zeng, Ailing, Liu, Minhao, Qin, Jing, & Xu, Qiang.

2018. Structure-Aware 3D Hourglass Network for Hand Pose Estimation from Single Depth Image. *CoRR*, **abs/1812.10320**.
- [8] Iqbal, Umar, Doering, Andreas, Yasin, Hashim, Krüger, Björn, Weber, Andreas, & Gall, Juergen. 2017. A Dual-Source Approach for 3D Human Pose Estimation from a Single Image. *CoRR*, **abs/1705.02883**.
- [9] Jiu, Mingyuan, Wolf, Christian, Taylor, Graham, & Baskurt, Atilla. 2014. Human Body Part Estimation from Depth Images via Spatially-Constrained Deep Learning. *Pattern Recogn. Lett.*, **50**(C), 122–129.
- [10] Joo, Hanbyul, Simon, Tomas, Li, Xulong, Liu, Hao, Tan, Lei, Gui, Lin, Banerjee, Sean, Godisart, Timothy Scott, Nabbe, Bart, Matthews, Iain, Kanade, Takeo, Nobuhara, Shohei, & Sheikh, Yaser. 2017. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [11] Lekhwani, Rohan. 2019. FastV2C-HandNet: Fast Voxel to Coordinate Hand Pose Estimation with 3D Convolutional Neural Networks. *ArXiv*, **abs/1907.06327**.
- [12] Malik, Jameel, Elhayek, Ahmed, & Stricker, Didier. 2018. Structure-Aware 3D Hand Pose Regression from a Single Depth Image. In: *EuroVR*.
- [13] Mann, H. B., & Whitney, D. R. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.*, **18**(1), 50–60.
- [14] Marin-Jimenez, M.J., Romero-Ramirez, F.J., Muñoz Salinas, R., & Medina-Carnincer, R. 2018. 3D Pose Estimation from Depth Maps using a Deep combination of Poses. *Journal of Visual Communication and Image Representation*. In press.

- [15] Martinez, Julieta, Hossain, Rayat, Romero, Javier, & Little, James J. 2017. A simple yet effective baseline for 3d human pose estimation. *CoRR*, **abs/1705.03098**.
- [16] Mehta, Dushyant, Rhodin, Helge, Casas, Dan, Fua, Pascal, Sotnychenko, Oleksandr, Xu, Weipeng, & Theobalt, Christian. 2017a. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *In: 3D Vision (3DV), 2017 Fifth International Conference on.* IEEE.
- [17] Mehta, Dushyant, Sridhar, Srinath, Sotnychenko, Oleksandr, Rhodin, Helge, Shafiei, Mohammad, Seidel, Hans-Peter, Xu, Weipeng, Casas, Dan, & Theobalt, Christian. 2017b (July). VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. vol. 36.
- [18] Mehta, Dushyant, Sotnychenko, Oleksandr, Mueller, Franziska, Xu, Weipeng, Sridhar, Srinath, Pons-Moll, Gerard, & Theobalt, Christian. 2018 (sep). Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB. *In: 3D Vision (3DV), 2018 Sixth International Conference on.* IEEE.
- [19] Mehta, Dushyant, Sotnychenko, Oleksandr, Mueller, Franziska, Xu, Weipeng, Elgharib, Mohamed, Fua, Pascal, Seidel, Hans-Peter, Rhodin, Helge, Pons-Moll, Gerard, & Theobalt, Christian. 2019. XNect: Real-time Multi-person 3D Human Pose Estimation with a Single RGB Camera. *CoRR*, **abs/1907.00837**.
- [20] Moon, Gyeongsik, Chang, Ju Yong, & Lee, Kyoung Mu. 2017. V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand

- and Human Pose Estimation from a Single Depth Map. *CoRR*, **abs/1711.07399**.
- [21] Moreno-Noguer, Francesc. 2016. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. *CoRR*, **abs/1611.09010**.
- [22] Newell, Alejandro, Yang, Kaiyu, & Deng, Jia. 2016. Stacked Hourglass Networks for Human Pose Estimation. *CoRR*, **abs/1603.06937**.
- [23] Photoneo. 2019. *3D Camera: MotionCam*. <https://www.photoneo.com/motioncam-3d/>. Accessed: 2020-03-27.
- [24] Qi, Charles Ruizhongtai, Su, Hao, Mo, Kaichun, & Guibas, Leonidas J. 2016. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *CoRR*, **abs/1612.00593**.
- [25] Qi, Charles Ruizhongtai, Yi, Li, Su, Hao, & Guibas, Leonidas J. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *CoRR*, **abs/1706.02413**.
- [26] Ren, Shaoqing, He, Kaiming, Girshick, Ross, & Sun, Jian. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Pages 91–99 of:* Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., & Garnett, R. (eds), *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc.
- [27] Rogez, Grégory, Weinzaepfel, Philippe, & Schmid, Cordelia. 2019. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 01.
- [28] Schnürer, Thomas, Fuchs, Stefan, Eisenbach, Markus, & Gross, Horst-

- Michael. 2019 (01). Real-time 3D Pose Estimation from Single Depth Images.
- [29] Shafaei, Alireza, & Little, James J. 2016. Real-Time Human Motion Capture with Multiple Depth Cameras. *In: Proceedings of the 13th Conference on Computer and Robot Vision*. Canadian Image Processing and Pattern Recognition Society (CIPPRS).
- [30] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. 2011 (June). Real-time human pose recognition in parts from single depth images. *Pages 1297–1304 of: CVPR 2011*.
- [31] Sun, Ke, Xiao, Bin, Liu, Dong, & Wang, Jingdong. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *CoRR, abs/1902.09212*.
- [32] Sun, Xiao, Shang, Jiaxiang, Liang, Shuang, & Wei, Yichen. 2017. Compositional Human Pose Regression. *CoRR, abs/1704.00159*.
- [33] Vidal, Rene, Bajcsy, Ruzena, Ofli, Ferda, Chaudhry, Rizwan, & Kurillo, Gregorij. 2013. Berkeley MHAD: A Comprehensive Multimodal Human Action Database. *Pages 53–60 of: Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV)*. WACV '13. Washington, DC, USA: IEEE Computer Society.
- [34] Wu, Wenxuan, Qi, Zhongang, & Li, Fuxin. 2018. PointConv: Deep Convolutional Networks on 3D Point Clouds. *CoRR, abs/1811.07246*.
- [35] Xiong, Fu, Zhang, Boshen, Xiao, Yang, Cao, Zhiguo, Yu, Taidong, Zhou, Joey Tianyi, & Yuan, Junsong. 2019 (October). A2J: Anchor-to-Joint

Regression Network for 3D Articulated Pose Estimation From a Single Depth Image. *In: The IEEE International Conference on Computer Vision (ICCV)*.

[36] Ye, M., Xianwang Wang, Yang, R., Liu Ren, & Pollefeys, M. 2011 (Nov).

Accurate 3D pose estimation from a single depth image. *Pages 731–738 of: 2011 International Conference on Computer Vision.*