Computing and Informatics, Vol. 40, 2021, 387–407, doi: 10.31577/cai_2021_2_387

HUMAN POSE ESTIMATION USING PER-POINT BODY REGION ASSIGNMENT

Dana Škorvánková, Martin Madaras

Faculty of Mathematics, Physics and Informatics
Comenius University
Mlynská dolina F1, 842 48 Bratislava, Slovakia
e-mail: {dana.skorvankova, madaras}@fmph.uniba.sk

Abstract. In recent years, the task of human pose estimation has become increasingly important, due to the large scale of usage, including VR applications, as well as higher-level tasks, such as human behavior understanding. In this paper, we introduce a novel two-stage deep learning approach named Segmentation-Guided Pose Estimation (SGPE). The pipeline is based on two neural networks working in a sequential fashion, while both models effectively process unorganized point clouds on the input. First, the segmentation network performs a pointwise classification into the corresponding body regions. In the next step, the point cloud with the per-point region assignment, forming the fourth input channel, is passed to the regression network. This way, both local and global features of the point cloud are preserved, helping the model fully maintain the body pose structure. Our strategy achieves competitive results on all of the examined benchmark datasets, and outperforms state-of-the-art methods.

Keywords: Machine vision, deep learning, neural networks, pose estimation, point clouds

Mathematics Subject Classification 2010: 68-T45

1 INTRODUCTION

One of many fields where the neural networks are applicable is the 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. 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, ergonomic body posture analysis, action recognition, 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.

Analyzing previous human pose estimation methods based on deep learning, the pipeline is usually formed by passing a single 2D image to the network, which directly regresses the 3D skeletal joint coordinates. Single-person pose estimation forms a basis for a number of related tasks, such as multi-person pose estimation [3, 18, 28], pose tracking [38, 40] or video pose estimation [24]. Most of the research is currently focused on estimating the pose from RGB data [2, 18, 25, 28, 33], mainly due to easily obtainable data which can be captured using a conventional RGB camera, without requiring a special hardware setup for recording. On the other hand, methods processing depth data on the input proved to be beneficial in terms of accuracy, by providing the additional spatial information.

Since most of the research is currently focused on estimating the pose from RGB data [5, 17, 18, 20, 23, 28], 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. Currently, there is a very limited 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, 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 usual workaround is to use the Kinect camera for recording, which can also directly extract the 3D skeleton joint coordinates, even though still working well only in indoor scenes.

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 [14, 15, 30, 39], thus encoding the third dimension into the 2D image. The depth maps are the 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 [21]. In an attempt to overcome these drawbacks, voxelized grids have been used in several solutions [9, 13, 21] 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.

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 [37]. 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.

Our work solves the task of single-person human pose estimation from depth data using a novel two-stage deep learning method called Segmentation-Guided Pose Estimation (SGPE). To avoid the projection of 3D human pose to 2D image space, we employ unorganized and unordered point clouds on the input to compute 3D skeletal joint coordinates as a result. We enhance the local and global feature propagation by performing an auxiliary semantic segmentation into the body regions. First, a corresponding body region is assigned to each point of the point cloud in a segmentation stage. To enable the network to fully perceive the data in its local as well as global context, we also make use of the intermediate concatenation of pointwise and aggregated features inside the model. Second, the input point cloud containing the point coordinates is concatenated with the per-point body region labels, adding the fourth channel to the data. Afterwards, the four-channel point clouds are fed into the regression model, which is where the resulting joint coordinates are estimated as an output. The main contributions of this work can be summarized as follows.

- We cope with the excessive number of network parameters and computational cost by processing depth data in a form of sparse unordered point clouds, instead of the commonly used depth maps. This way, we also avoid the need for the model to perform a distortion-invariant estimation.

- Our two-stage pipeline deals with the issues related to poor propagation of local context through the networks, by concatenation of features extracted in intermediate layers before and after pooling aggregation, and by incorporating residual connections in-between the layers. Thus, we improve the gradient flow inside the models. Furthermore, to increase the accuracy of estimated joint coordinates, we augment the initial 3D point clouds with a per-point body region segmentation predicted in the first stage of the pipeline.
- To evaluate our approach, we conduct experiments on a number of depth-based human pose benchmark datasets, including both synthetic and real data. Our strategy achieves competitive results on all of the examined datasets, and outperforms state-of-the-art methods.

2 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 [36], 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) [5, 17, 18, 20, 23, 28, 33, 42], and three-dimensional data (depth maps, point clouds, voxelized grids etc.) [1, 6, 7, 8, 11, 15, 21, 31, 32, 39, 41]. 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, which 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.

2.1 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 [16, 34] or use the 2D pose to infer the 3D pose [4, 17, 19, 20, 28]. Among those employing the 2D pose, many approaches make use of lifting the estimated 2D pose to 3D [4, 10, 16, 22] by direct regression, database matching etc.

One of the first real-time approaches was proposed by Mehta et al. [20]. They introduced a system 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. However, the stated model was unable to handle occlusions. Thus, they removed the restrictions in the follow-up work [18], where the model is also extended 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.

2.2 Depth-Based Human Pose Estimation

The depth data used as the input to the neural networks comes in various forms. Most frequently, the 3D input data is in a form of a depth map (RGB-D image). 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. [15] 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. 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 and Little [31]. 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.

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. However, voxels require use of three-dimensional convolutions, which makes operations with them very time-consuming and computationally expensive. *V2V PoseNet* [21] 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. 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.

2.3 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. [26] 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. [27] introduced *PointNet++* model, which has similar key structure as the previous PointNet, but it 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.

In another study, Wu et al. [37] 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.

It is worth mentioning, that all of the stated methods processing unordered point clouds perform object classification or segmentation task, which is not an aim of this work. 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, compared to 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 estimations.

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.

In this paper, we solve the problem of depth-based human pose estimation using unordered point clouds as the input data type. However, unlike the previous approaches processing point clouds, our pipeline works in two subsequent stages, instead of a direct regression, to effectively merge both local and global features of the data without losing any contextual information. Thus, the resulting pose coordinates can be regressed from a point cloud enhanced by additional regional information, helping the network fully maintain the body pose structure.

3 OVERVIEW

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, as depicted in Figure 1, makes use of residual connections added to the shared multi-layer perceptron blocks, to strengthen the feature propagation.

4 SEGMENTATION-GUIDED POSE ESTIMATION

This section describes our proposed method in detail, providing further information on the training procedures. Our pipeline takes a point cloud on the input, passes it through two subsequent neural networks, and outputs the 3D coordinates of the skeletal joints, defining the estimated human pose. 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 $p = 2\,048$, 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

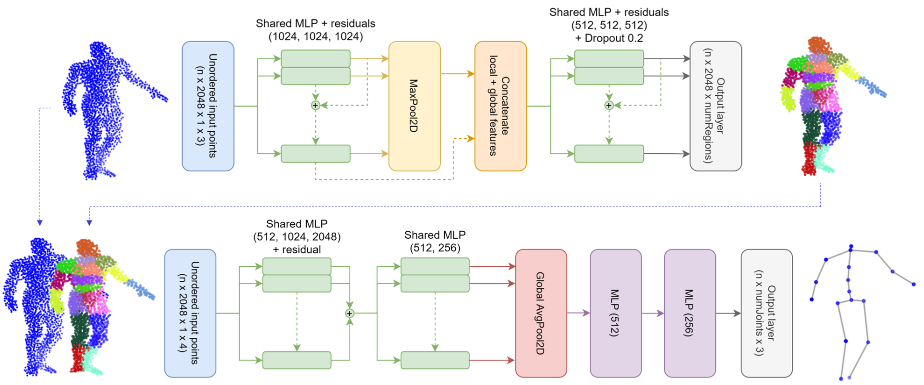

Figure 1. The overview of the proposed Segmentation-Guided Pose Estimation pipeline: First, the point clouds are segmented into body regions in the segmentation network (*top*), then the input point clouds are concatenated with the predicted per-point body region assignment as a fourth channel, and fed into the regression network (*bottom*)

$[-1, 1]$ along each axis, using minimum and maximum values of the whole training set.

4.1 Shared Multi-Layer Perceptron Module

The shared multi-layer perceptron (MLP), introduced in [26], is a stack of convolutional layers with kernel size 1×1. Unlike the standard convolutional layers, they do not affect the dimension of the input, but instead expand (or shrink) the dimension of the filters. By operating in the filter space, the 1×1 convolutions allow us to process unorganized and permutation-invariant sets of points. The points passed to the shared MLP module are treated as 2D input with dimensions 1×3.

4.2 Body Region Segmentation Network

As a 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 (e.g. U-net [29] or PointNet [26]). Instead of using an exhausting segmentation architecture, which has high memory and time requirements, we decided to utilize the same main modules in the segmentation and regression model. This is partly because segmentation is not the main task of this work, and is strictly in role of an auxiliary subtask, therefore the absolute segmentation accuracy is not crucial in our study. Also, 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 1 (top), is based on the shared multi-layer perceptron 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. The number of regions matches the number of joints in skeleton, each region being associated with the particular joint. Every single point of the point cloud is then assigned to the region corresponding to the nearest skeleton node in terms of Euclidean distance.

4.3 Regression Network

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 1, 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. Finally, the model estimates the 3D skeletal joint coordinates of the captured human subject as the output.

5 RESULTS

5.1 Benchmark Datasets

ITOP. The ITOP dataset [8] contains 40 K training and 10 K testing depth frames recorded from two viewpoints (front-view and top-view). The dataset captures 20 different subjects, each performing 15 sequences. The ground truth skeleton is defined by 3D coordinates of 15 skeletal joints.

UBC3V. The UBC3V [31] is a synthetically made human pose dataset. It contains around 6 M synthetic depth frames structured in three parts according to the complexity of the human postures – easy, medium and hard pose, each with its train, validation and test split. 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.

MHAD. The MHAD dataset [35] consists of 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 250 K. The skeleton structure in this dataset contains 35 joints.

CMU Panoptic dataset. The CMU Panoptic [12] is a large scale multi-modal human pose 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 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 skeleton structure consists of 15 joint locations. The database captures multiple actors of different gender, age and body shape.

5.2 Evaluation Metrics

In the process of evaluation, we used mean per joint position error (MPJPE) and mean average precision (mAP) as metrics, following [1, 8, 15, 31]. Mean average precision is defined as percentage of all skeletal joints predicted under 10 cm threshold from ground truth.

5.3 Implementation Details

We conduct experiments on NVIDIA GTX 1070. 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.

Regarding 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, mean absolute error between the predicted locations and the ground truth labels of all skeletal joints is used to determine the model loss. We have also evaluated the performance of the regression network using huber loss with a regularization term, yielding approximately the same estimation accuracy.

For the regularization purposes, a single dropout layer with rate of 0.2 is included before the output layer of the segmentation network (as shown in Figure 1 (top)).

5.4 Experiments

For the purpose of evaluation, we used several benchmark datasets, including the challenging ITOP front-view [8], UBC3V hard-pose [31], MHAD [35] and a subset of CMU Panoptic dataset [12]. 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 5, left). Using a 10 cm threshold, the mean average precision is 85.57 %, which is comparable to the state-of-the-art results.

Regarding the CMU dataset, we evaluated our method specifically on the *Range of motion* section of the dataset, yielding approximately 141 K 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 5, right), and the mean average precision at 10 cm is 98.39 %. Figure 2 illustrates the qualitative results on samples from CMU Panoptic dataset.

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

1. choosing the test set as randomly sampled 25 % of the dataset,
2. leave-one-subject-out cross validation.

In case of MHAD data, the original skeleton is rather complex, containing as many as 35 skeleton nodes. 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 3), in the same way as in [1]. However, we present

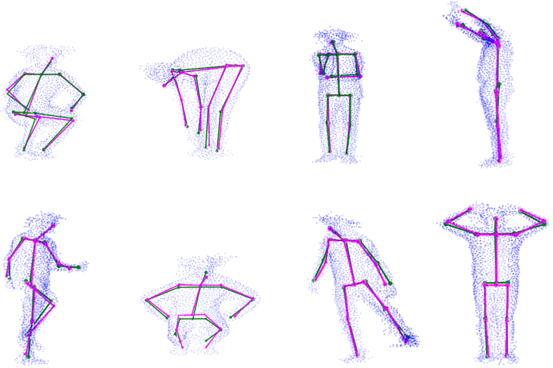

Figure 2. Qualitative results of our method on CMU Panoptic dataset [12]. The ground truth skeletons (*green*) vs. our estimation (*magenta*). Best viewed in color.

results of our approach also on the original full skeleton, to be able to compare our strategy to the existing methods (as shown in Table 1). 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. As it can be seen in Table 1, the mean per joint position error has visibly decreased after omitting the redundant skeletal joints.

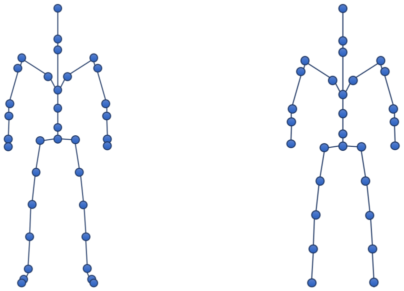

Figure 3. The original skeleton structure used in MHAD dataset (*left*) vs. the modified skeleton (*right*)

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 4, 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 6). We set a novel state-of-

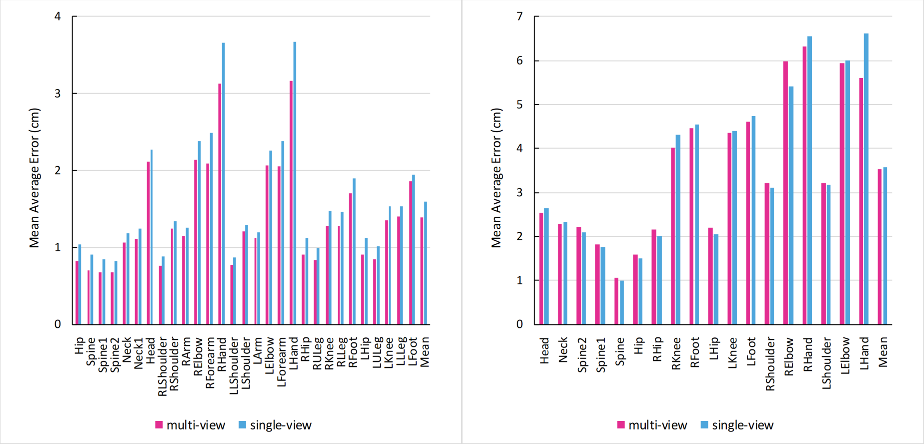

Figure 4. The mean per joint position error (MPJPE) on MHAD (*left*) and UBC3V (*right*) datasets, comparing multi-view and single-view approach

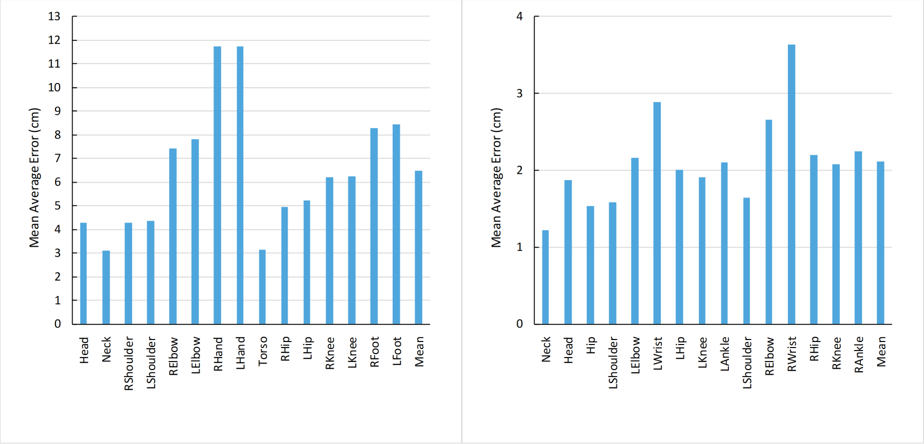

Figure 5. The mean per joint position error (MPJPE) on ITOP (*left*) and CMU (*right*) datasets

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 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 3.36 cm in the case of single-view data, and 3.53 cm with multi-view data (as shown in Figure 4, right). The mean average precision at 10 cm is 95.63 % and 95.71 % for the single-view and multi-view

Method	Eval. Protocol	MPJPE [cm] Single-View	MPJPE [cm] Multi-View
Shafei et. al [31]	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**

Table 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

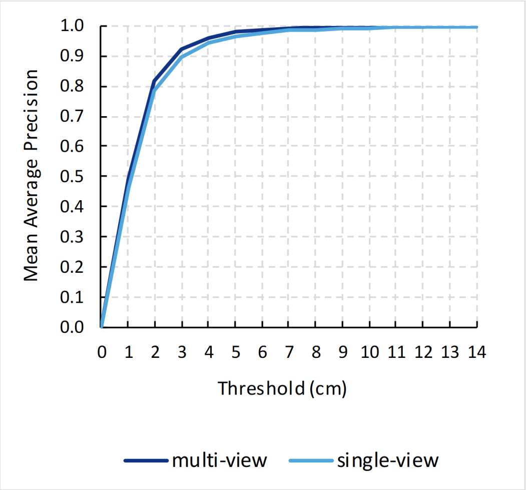

Figure 6. Mean average precision at 10 cm threshold on MHAD dataset for multi-view and single-view approaches

approach respectively. The claimed results of the Deep Depth Pose (DDP) model proposed in [15] are listed in italics, due to a number of unsuccessful attemps 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, predicted on merged multi-view point clouds.

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 8. Our method achieves up to 95 % segmentation accuracy on CMU Panoptic dataset.

Method	MPJPE (cm) Single-View	MPJPE (cm) Multi-View
DDP (observed)	19.23	–
PBPE [1]	7.59	5.59
Shafei et. al [31]	–	5.64
DDP (claimed) [15]	*3.15*	*2.36*
Ours – FCPE	**3.57**	**3.53**

Table 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

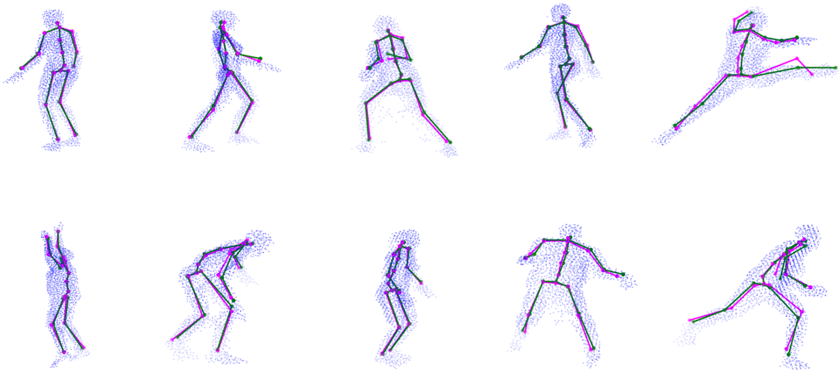

Figure 7. Qualitative results of our approach on test set of UBC hard-pose dataset [31]. The ground truth skeletons (*green*) vs. our estimation (*magenta*). Best viewed in color.

6 LIMITATIONS

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

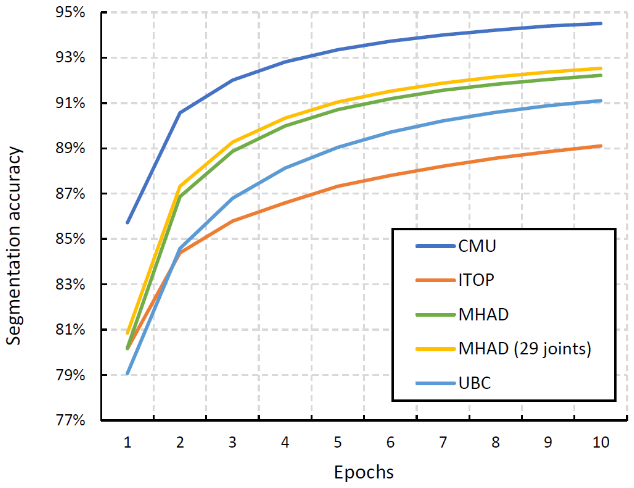

Figure 8. Accuracy of the body-parts segmentation performed in the first stage of our pipeline on all examined datasets

depth sensors need to be synchronized mutually as well as with the motion capture system.

7 CONCLUSIONS

We proposed a novel method for the accurate single-person depth-based human pose estimation called Segmentation-Guided Pose Estimation (SGPE). Main contribution of our work is the elimination of 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. 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.

Acknowledgement

Our research was supported from the grant No. UK/91/2021.

REFERENCES

[1] Ali, A.: 3D Human Pose Estimation. M.Sc. Thesis, Georgia Institute of Technology, May 2019.

[2] Artacho, B.—Savakis, A.: UniPose: Unified Human Pose Estimation in Single Images and Videos. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 7033–7042, doi: 10.1109/CVPR42600.2020.00706.

[3] Briq, R.—Doering, A.—Gall, J.: Unifying Part Detection and Association for Recurrent Multi-Person Pose Estimation. CoRR, 2019, arXiv: 1904.11864.

[4] Chen, C. H.—Ramanan, D.: 3D Human Pose Estimation = 2D Pose Estimation + Matching. CoRR, 2016, arXiv: 1612.06524.

[5] Chou, C. J.—Chien, J. T.—Chen, H. T.: Self Adversarial Training for Human Pose Estimation. CoRR, 2017, arXiv: 1707.02439.

[6] Ge, L.—Liang, H.—Yuan, J.—Thalmann, D.: 3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 5679–5688, doi: 10.1109/CVPR.2017.602.

[7] Ge, L.—Liang, H.—Yuan, J.—Thalmann, D.: Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 3593–3601, doi: 10.1109/cvpr.2016.391.

[8] Haque, A.—Peng, B.—Luo, Z.—Alahi, A.—Yeung, S.—Fei-Fei, L.: Towards Viewpoint Invariant 3D Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.): Computer Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9905, 2016, pp. 160–177, doi: 10.1007/978-3-319-46448-0_10.

[9] Huang, F.—Zeng, A.—Liu, M.—Qin, J.—Xu, Q.: Structure-Aware 3D Hourglass Network for Hand Pose Estimation from Single Depth Image. CoRR, 2018, arXiv: 1812.10320.

[10] Iqbal, U.—Doering, A.—Yasin, H.—Krüger, B.—Weber, A.—Gall, J.: A Dual-Source Approach for 3D Human Pose Estimation from a Single Image. CoRR, 2017, arXiv: 1705.02883.

[11] Jiu, M.—Wolf, C.—Taylor, G.—Baskurt, A.: Human Body Part Estimation from Depth Images via Spatially-Constrained Deep Learning. Pattern Recognition Letters, Vol. 50 (C), 2014, pp. 122–129, doi: 10.1016/j.patrec.2013.09.021.

[12] Joo, H.—Simon, T.—Li, X.—Liu, H.—Tan, L.—Gui, L.—Banerjee, S.—Godisart, T. S.—Nabbe, B.—Matthews, I.—Kanade, T.—Nobuhara, S.—Sheikh, Y.: Panoptic Studio: A Massively Multiview System for Social Interaction

Capture. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, 2017, No. 1, pp. 190–204, doi: 10.1109/TPAMI.2017.2782743.

[13] Lekhwani, R.—Singh, B.: FastV2C-HandNet: Fast Voxel to Coordinate Hand Pose Estimation with 3D Convolutional Neural Networks. CoRR, 2019, arXiv: 1907.06327.

[14] Malik, J.—Elhayek, A.—Stricker, D.: Structure-Aware 3D Hand Pose Regression from a Single Depth Image. In: Bourdot, P., Cobb, S., Interrante, V., Kato, H., Stricker, D. (Eds.): Virtual Reality and Augmented Reality (EuroVR 2018). Springer, Cham, Lecture Notes in Computer Science, Vol. 11162, 2018, pp. 3–17, doi: 10.1007/978-3-030-01790-3_1.

[15] Marín-Jiménez, M. J.—Romero-Ramirez, F. J.—Muñoz-Salinas, R.—Medina-Carnicer, R.: 3D Human Pose Estimation from Depth Maps Using a Deep Combination of Poses. Journal of Visual Communication and Image Representation, Vol. 55, 2018, pp. 627–639, doi: 10.1016/j.jvcir.2018.07.010.

[16] Martinez, J.—Hossain, R.—Romero, J.—Little, J. J.: A Simple Yet Effective Baseline for 3D Human Pose Estimation. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2659–2668, doi: 10.1109/iccv.2017.288.

[17] Mehta, D.—Rhodin, H.—Casas, D.—Fua, P.—Sotnychenko, O.—Xu, W.—Theobalt, C.: Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. 2017 International Conference on 3D Vision (3DV), Qingdao, China, IEEE, 2017, pp. 506–516, doi: 10.1109/3dv.2017.00064.

[18] Mehta, D.—Sotnychenko, O.—Mueller, F.—Xu, W.—Elgharib, M.—Fua, P.—Seidel, H.-P.—Rhodin, H.—Pons-Moll, G.—Theobalt, C.: XNect: Real-Time Multi-Person 3D Human Pose Estimation with a Single RGB Camera. CoRR, 2019, arXiv: 1907.00837v1.

[19] Mehta, D.—Sotnychenko, O.—Mueller, F.—Xu, W.—Sridhar, S.—Pons-Moll, G.—Theobalt, C.: Single-Shot Multi-Person 3D Pose Estimation from Monocular RGB. 2018 International Conference on 3D Vision (3DV), Verona, Italy, IEEE, 2018, pp. 120–130, doi: 10.1109/3dv.2018.00024.

[20] Mehta, D.—Sridhar, S.—Sotnychenko, O.—Rhodin, H.—Shafiei, M.—Seidel, H.-P.—Xu, W.—Casas, D.—Theobalt, C.: VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera. ACM Transactions on Graphics, Vol. 36, 2017, No. 4, Art. No. 44, 14 pp., doi: 10.1145/3072959.3073596.

[21] Moon, G.—Chang, J. Y.—Lee, K. M.: V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 5079–5088, doi: 10.1109/CVPR.2018.00533.

[22] Moreno-Noguer, F.: 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1561–1570, doi: 10.1109/CVPR.2017.170.

[23] Newell, A.—Yang, K.—Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.): Computer Vision – ECCV 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9912, 2016, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.

[24] Pavllo, D.—Feichtenhofer, C.—Grangier, D.—Auli, M.: 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 7745–7754, doi: 10.1109/CVPR.2019.00794.

[25] Peng, X.—Tang, Z.—Yang, F.—Feris, R. S.—Metaxas, D. N.: Jointly Optimize Data Augmentation and Network Training: Adversarial Data Augmentation in Human Pose Estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 2226–2234, doi: 10.1109/cvpr.2018.00237.

[26] Qi, C. R.—Su, H.—Mo, K.—Guibas, L. J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[27] Qi, C. R.—Yi, L.—Su, H.—Guibas, L. J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. CoRR, 2017, arXiv: 1706.02413.

[28] Rogez, G.—Weinzaepfel, P.—Schmid, C.: LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, 2020, No. 5, pp. 1146–1161, doi: 10.1109/TPAMI.2019.2892985.

[29] Ronneberger, O.—Fischer, P.—Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (Eds.): Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer, Cham, Lecture Notes in Computer Science, Vol. 9351, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[30] Schnürer, T.—Fuchs, S.—Eisenbach, M.—Gross, H.-M.: Real-Time 3D Pose Estimation from Single Depth Images. Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Volume 5: VISAPP, Prague, Czech Republic, 2019, pp. 716–724, doi: 10.5220/0007394707160724.

[31] Shafaei, A.—Little, J. J.: Real-Time Human Motion Capture with Multiple Depth Cameras. 2016 13th Conference on Computer and Robot Vision (CRV), Victoria, BC, Canada, 2016, pp. 24–31, doi: 10.1109/CRV.2016.25.

[32] Shotton, J.—Fitzgibbon, A.—Cook, M.—Sharp, T.—Finocchio, M.—Moore, R.—Kipman, A.—Blake, A.: Real-Time Human Pose Recognition in Parts from Single Depth Images. 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO, USA, 2011, pp. 1297–1304, doi: 10.1109/CVPR.2011.5995316.

[33] Sun, K.—Xiao, B.—Liu, D.—Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5686–5696, doi: 10.1109/cvpr.2019.00584.

[34] Sun, X.—Shang, J.—Liang, S.—Wei, Y.: Compositional Human Pose Regression. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2621–2630, doi: 10.1109/iccv.2017.284.

[35] Ofli, F.—Chaudhry, R.—Kurillo, G.—Vidal, R.—Bajcsy, R.: Berkeley MHAD: A Comprehensive Multimodal Human Action Database. 2013 IEEE Workshop on Applications of Computer Vision (WACV), 2013, pp. 53–60, Clearwater Beach, FL, USA, doi: 10.1109/WACV.2013.6474999.

[36] Wang, P.—Li, W.—Ogunbona, P.—Wan, J.—Escalera, S.: RGB-D-Based Human Motion Recognition with Deep Learning: A Survey. CoRR, 2017, arXiv: 1711.08362.

[37] Wu, W.—Qi, Z.—Li, F.: PointConv: Deep Convolutional Networks on 3D Point Clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9613–9622, doi: 10.1109/CVPR.2019.00985.

[38] Xiao, B.—Wu, H.—Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11210, 2018, pp. 472–487, doi: 10.1007/978-3-030-01231-1_29.

[39] Xiong, F.—Zhang, B.—Xiao, Y.—Cao, Z.—Yu, T.—Zhou, J. T.—Yuan, J.: A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 2019, pp. 793–802, doi: 10.1109/ICCV.2019.00088.

[40] Xiu, Y.—Li, J.—Wang, H.—Fang, Y.—Lu, C.: Pose Flow: Efficient Online Pose Tracking. 29th British Machine Vision Conference (BMVC), Newcastle, UK, CoRR, 2018, arXiv: 1802.00977.

[41] Ye, M.—Wang, X.—Yang, R.—Ren, L.—Pollefeys, M.: Accurate 3D Pose Estimation from a Single Depth Image. 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 731–738, doi: 10.1109/ICCV.2011.6126310.

[42] Yin, B.—Zhang, D.—Li, S.—Hao, A.—Qin, H.: Context-Aware Network for 3D Human Pose Estimation from Monocular RGB Image. 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1–8, doi: 10.1109/IJCNN.2019.8852263.

Dana Škorvánková is Ph.D. student at the Faculty of Mathemathics, Physics and Informatics of the Comenius University in Bratislava. Her dissertation thesis is oriented on neural networks applied on skeleton tracking and anthropometric measurements estimation. She finished her bachelor and master degree at the stated faculty, with focus on computer graphics and computer vision, along with machine learning and neural networks. The bachelor thesis "Capturing of Movement During Music Performance" was selected as the best thesis of the Department of Applied Informatics, and the master thesis "Deep Learning-Based Human Pose Estimation from 3D Data" has been awarded by the head of the university.

Martin Madaras received his Ph.D. degree in computer science in 2014 from the Comenius University in Bratislava with the focus on mesh processing and skeleton applications. In 2017 he co-founded a company Skeletex Research, where the main focus has been given on processing data from 3D scanners and cameras. He has been Postdoctoral Researcher at the Comenius University for 6 years and research lead at Skeletex Research for 4 years. Currently, he is also working as Assistant Professor at the Slovak University of Technology. The main research topics in his scope are geometry and point cloud processing, skeleton tracking and 3D model reconstruction.