Estimating Human pose from depth images using Convolutional Neural Networks

Both Eyes Open

Bård-Kristian Krohg

Thesis submitted for the degree of
Master in Informatics: Robotics and Intelligent Systems
60 credits

Institute for informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2021

Estimating Human pose from depth images using Convolutional Neural Networks

Both Eyes Open

Bård-Kristian Krohg

Estimating Human pose from depth images using Convolutional Neural Networks

http://www.duo.uio.no/

Printed: X-press printing house

Estimating Human pose from depth images using Convolutional Neural Networks

Bård-Kristian Krohg

1st June 2021

Abstract

This work is part of a larger project that explores the possibility of bringing robotics into geriatric care. The goal of that project is to create a robotic system that can assist in optimizing the use of caregivers, so they are used where they are needed.

This work introduces *DepthPose* – a system that solves human pose estimation in 3D for full skeletons. The system is lightweight and can be deployed on mobile units.

Convolutional Neural Networks have been used for solving object recognition in 2D images with great success. This work aims to use the same techniques to extract 3D human poses from depth images in real-time. Two multi-staged CNNs, one to encode the location of each joint and another to encode the association between the joints, provide initial poses and locations for perceived people. The locations are then refined using a novel articulation network.

Preface

First, I would like to thank my supervisors Jim Tørresen and Ryo Kurazume, for their support, guidance, and patience during the development of this project. Second, my HR manager Marit Flendstad Kruse, for her assistance in letting me combine work with the writing of this thesis. Last, I would like to thank everyone at the Kurazume-lab for their welcome and help during my stay at Kyushu University, and my friends and family for proofreading and encouragement.

The subtitle, Both Eyes Open, has a double meaning: In this paper, we will explore the world in 3D. The biological way to achieve depth vision is by using two eyes, hence both eyes open. The other way to interpret the subtitle is tied with the small figure on the front page. In Japanese, the saying translated as "Both Eyes Open" refers to the Daruma figure.

When one is working toward a goal, such as completing a thesis, one can purchase a Daruma figure from a temple. When bought, the Daruma has blank eyes - they are closed. The buyer then paints in one eye, asking for the Darumas help in completing their goal. In exchange for the Darumas help, the buyer promises to paint in the other eye. One significant detail about the Daruma is that it is weighted on the bottom. If it should ever falter and fall over, it will right itself back up and continue on its way to completing the goal.

The ability to right yourself up for every setback has been of particular inspiration to me during my work on this thesis. This is why I have placed the figure on the front page; as a personal helper.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As life expectancy increases in Norway, so does the population who needs geriatric care either at home or in a geriatric facility. According to a communal health survey [17], it is projected that the increasing senior population in Oslo will lead to increased pressure on the healthcare services. To mitigate the need for help at home, it is important to encourage health-promoting activities and uncover what and where preventative measures are needed. The Multimodal Elderly Care System (MECS) is proposed as a solution where data gathered in users' homes guide where and what help is needed from healthcare personnel – be it preventative or acute.

The MECS is envisioned as a small, mobile unit that can be introduced to any home. This eliminates the user from the operational loop of the unit, making the data gathering process robust to forgetfulness or physical ability of the user. One of the key information points gathered by the MECS will be the users' physical pose, which informs the system of the following:
Body language, enabling the possibility of smooth Human-Robot Interaction (HRI) when moving around or determining the intent of the user.
Physical state, informing the MECS whether the user is in need of acute help.
History of natural posture, which could be invaluable information for a physical therapist or doctor to develop personalized training programs preventing muscle degradation.
Activity recognition, helping the MECS decide on what actions to execute. For example, reminding the user to take their prescribed

medication if they forget.

A 2D system could capture all these key points; however, a 2D representation of a pose would not be robust to different viewing angles. A 3D representation could define the origin of two poses' coordinate system to the same landmark in each pose, making it is easy to compare the two poses by, for example, the euclidian distance between the poses' corresponding landmarks.

Estimating human pose in 3D, or Motion Capture (mocap) is a well-known area of research. However, mocap is expensive and requires a large amount of physical hardware. The industry standard is to use an elaborate mocap studio that requires multiple expensive cameras, a large area, and specialized software. Therefore the application areas for motion capture are currently mostly limited to research, movie-, and videogame-making. The mocap problem deals with finding a representation of an actor, creature, or object that can be used in animation. This representation is often a *rigged* skeleton with bones that define the movement of the animated character [21]. The movements of the computerized rig are mapped to the movements of an observed actor in the real world and recorded.

1.1 Research Goals

This work explores the possibility of using a single depth camera to solve the mocap problem of estimating human pose in 3D. The resulting method needs to be lightweight and fast enough to be viable in a mobile unit with limited processing capability. The resulting method must also be accurate enough for healthcare personnel or the MECS itself to extract reliable and useful information about the pose.

The method presented in this thesis will therefore be evaluated using the following goals:

1. Propose a lightweight system for recognizing human pose in 3D based on depth images.

2. Design and develop an architecture that can be deployed to systems with limited hardware, and provide useful information.

3. Evaluate the performance of the system and find room for improvements for it.

1.2 Contributions

The main contribution of this work, a system called *DepthPose*, – is the proposal of using a shallow CNN trained on depth images, in combination with a novel articulation network to define human pose in 3D. The CNNs main task is to propose a set of poses present in the depth image. These poses are refined in the articulation network, which allows for a shallower CNN architecture. This leads then to fewer parameters in the system, which leads to a more lightweight method that should preform faster on limited hardware.

Chapter 2

Background

The classical approach to extract useful information from an image has been to find mathematical definitions for features that describe the information. The features could, for example, be lines, circles, or edges. Circles can be used to find coins, where the diameter denotes the value. Lines can be used to find how many fenceposts are in a fence. Traditional mathematical models include the Hough Transform [13] for detecting lines or circles, the Sobel operator to detect gradients, and in return, edges, or the Gray Level Co-occurrence Matrix for detecting texture features. In common for all these methods is that they are well defined and find precisely *one* type of feature that was previously specified. At a higher level, hand-crafted feature descriptors have been made. They look for a specific set of features in an image to identify unique locations. Some well-known examples are the SIFT [20], SURF [2] and ORB [25] feature descriptors. They have been used successfully in applications such as combining images into panoramas or combining different views to a 3D model, also known as Structure from motion [32].

Hand-crafted features have also been used in machine learning scenarios. As an example, Haar-like features were demonstrated to efficiently find faces in [33]. Here, the machine learning model decided which of a given set of features to use to classify whether the frame contained a face or not. In using both the Sobel operator and using Haar-like features, a filter is passed over the image. In the case of the Sobel operator, this is an actual convolution of the image with the filter, whereas Haar-features are extracted using a sliding window. This is the intuition that is used in Convolutional Neural Networks, discussed next.

2.1 Convolutional Neural Networks

First introduced by Fukushima [9] CNNs encode spatial information without being affected by shifts in the position of that information. In other words, they look at collections of *spacially connected pixels*. CNNs use filters, or kernels, to extract features that help them to "understand" an image. The deeper the network is, the more complex features emerge. In addition, the *receptive field*, the patch of the original image that affects the feature, becomes larger.

Instead of using hand-crafted features, deep CNNs define the features they use when they are trained. This is one of the reasons why CNNs needs fairly large training sets, as well as taking a long time to train. Therefore, a popular approach is to reuse the first few primitive layers (the first layers encountered in a forward-pass) from other models. This is known as *Transfer Learning*. It can be accomplished by first training a neural network for a specific task, for example, recognizing a handful of different types of objects in an image. Then, the primitive layers with their parameters are "chopped off" the network. Since these layers only have learned primitive features, these learned features can be input and fine-tuned to specialize a different network, which in turn needs less training time since it has already learned the simple features. Another advantage of this is that the two networks can be trained on vastly different datasets. If one application area has an abundance of training data, the first network could be trained on this. As long as the general input is the same, namely the number of channels and the values presented, the first few layers can be easily fine-tuned to a new application area.

When designing a CNN, it is often helpful to have in mind exactly what one can imagine should be detected in each layer. In 2D images, the first few layers of a deep CNN have been shown to often mimic the behavior of the simple handmade feature extractors such as the ones discussed above. However, such features can be quite different in 3D. Instead of edge-detectors, one can imagine plane detectors or other detectors unique to 3D objects. Additionally, 3D depth images only have a single channel, whereas most 2D CNNs have been trained on 3-channel RGB images. [29] argues that it is, therefore, better to train such networks from scratch.

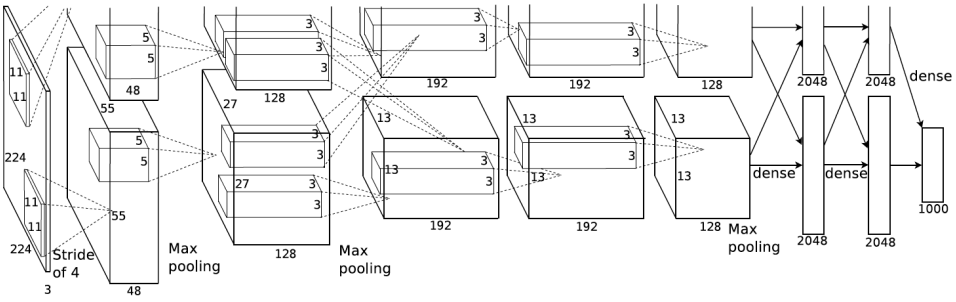


Figure 2.1: Illustration of a Convolutional Neural Network. The figure is borrowed from AlexNet [18]

2.2 Depth Images

In contrast to color, or RGB images, depth images contain information about the distances. Each pixel contains a measure of the distance from the camera x,y plane to the real world point projected through that pixel.

2.3 Pose Estimation

Human Pose Estimation is an area well-researched area because of its many applications. Much of the research focuses on finding pose in RGB images, where hand-crafted features such as parallel edges [22, 24], silhouette features [11], or pictorial structures [7, 8], have been used to suggest candidates for limbs, joints, and so on. From such evidence, constrained kinematic models, tree graphs, or decision trees [27], are used to construct the poses with varying degrees of success. With such top-down approaches, will the runtime of the algorithm increase when multiple people are present in the image.

With the increasing popularity of CNNs in recent years, many techniques utilizing them have also been developed for human pose estimation [31]. A particularly popular approach is to use a deep CNN to produce heat-maps that suggest candidate locations for various body parts [34, 35].

Much research has been done in estimating human pose in two dimensions, as quite large datasets have been made such, as the MPII, or the Human 3.6M datasets [1, 14].

The other approach is to use object recognition to find key features for

the whole image. Then the recognized landmarks are combined to build up the people instances.

In [4], the second approach is used. Two CNNs, one for landmark localization and the other for recombination, are trained to estimate human pose. One of the networks produces a *confidence map* for each joint. Each pixel in the confidence map contains the probability, or confidence, that the pixel is part of a person's joint. The other creates Part Affinity Field (PAF)s, which is a map of vectors pointing in the direction of one of M limbs. [1] These maps are assembled by bipartite matching to create the 2D skeletons observed in the scene.

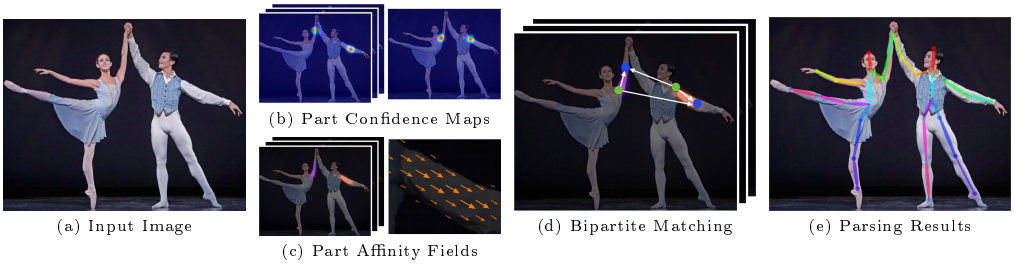


Figure 2.2: The pipeline used in OpenPose [3, 4]. The input image 2.2a is fed into the two networks, which produce joint detections in confidence maps 2.2b and PAFs 2.2c. Bipartite matching is performed in 2.2d, to determine which detected joints should be connected by a limb. 2.2e shows the finished results.

Using the method described in [4], no information is given for joints that are not accurately detected. If, for example, an extremity is occluded together with half of the connecting limb, the extremity will not be part of the output skeleton, even if the joint could be extrapolated from the parts of the limb that is visible. This is also true for undetected joints in the middle of a joint chain. The joint could be extrapolated using the surrounding joints. The problem with joint-extrapolation happens because of the bipartite matching, which does not work if any joint is missing.

[1] This work will stick to the convention of using the term *limb* to describe any *connection between any pair of body landmarks*. The body landmarks will be termed *joints*.

Chapter 3

Datasets

This chapter presents an overview of the datasets suitable for human pose estimation in depth images. The datasets are evaluated according to closeness to real-world conditions, the amount of data, and the accuracy of the recorded poses.

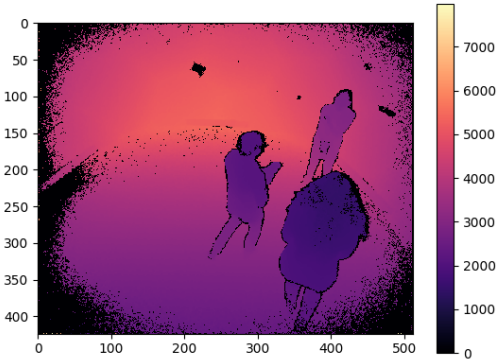
3.1 The Panoptic Studio Dataset

The dataset used in this work was created from source material provided by the Panoptic Studio [15, 16]. Calibration files, predicted ground-truth skeletons, and depth images were downloaded using the GNU parallel program [30]. Some problems were experienced, as the Panoptic Studio server was quite unreliable, which resulted in significant delays, as well as corrupted files and, in return, a smaller dataset than desired.

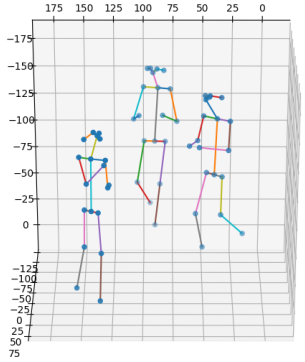
A set of five complete sequences were set aside as the test set to ensure that the network would not recognize any similar frames or body positions from previous sets.

The dataset contains samples from a wide variety of activities, view-angles, the number of people in the scene, and body compositions.

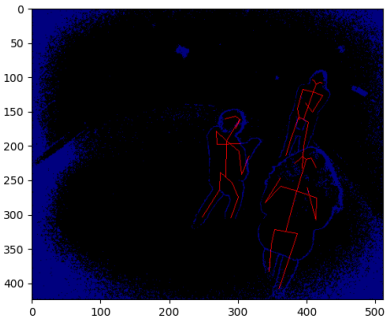
The sequences in the dataset were recorded at approximately 30 fps. All frames and world coordinate positions at an approximate single timestep are hereby referred to as a *frame*. To allow for poses to change, two samples were extracted per second. Each sequence was recorded by ten different Kinects from a height of 1 and 2 meters above the ground. This results in a total number of 10x the amount of samples listed in Table 3.1.



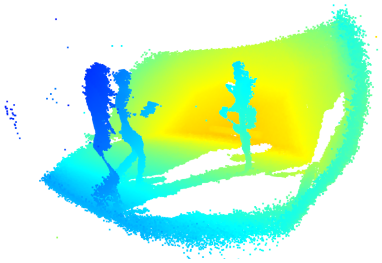
(a) Depth image from `KINECTNODE6`.



(b) Skeletons defined in the world coordinate system.



(c) Skeletons reprojected to the depth image



(d) The depth image projected to a point cloud.

Figure 3.1: Combination of datapoints for sequence 160226_haggling1, frame 144 in the Panoptic Studio Dataset [16].

	Category	Sequence Name	Duration (MM:SS)
Training / Validation	Range of Motion	171204_pose1	17:30
	Range of Motion	171204_pose2	22:30
	Range of Motion	171204_pose3	5:00
	Range of Motion	171204_pose5	15:00
	Range of Motion	171204_pose6	12:50
	Range of Motion	171026_pose1	13:20
	Range of Motion	171026_pose3	4:20
	Social Games	160226_haggling1	8:00
	Social Games	160422_haggling1	8:00
	Social Games	160422_ultimatum1	15:00
	Musical Instruments	160906_band1	1:00
	Musical Instruments	160906_band2	5:00
	Musical Instruments	160906_band3	5:00
	Toddler	160906_ian2	5:00
	Others	170915_office1	3:00
	Others	160906_pizza1	5:00
	Dance	170307_dance5	6:40
Test	Range of Motion	171204_pose4	17:30
	Range of Motion	171026_pose2	9:00
	Social Games	160224_haggling1	5:00
	Toddler	160906_ian1	2:00
	Others	170407_office2	3:00

Table 3.1: Sequences used for training, validation and testing

3.2 Dataset Augmentations

The coco19 points from the pose directories were reprojected into the depth image using the perspective camera model

$$\boldsymbol{u} = K \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R_{3\times3} & \boldsymbol{t}_{3\times1} \\ \boldsymbol{0}_{1\times3} & 1 \end{bmatrix} \tilde{\boldsymbol{X}} \tag{3.1}$$

The matrix K is the camera intrinsics and describes how to transform the point from the normalized image-plane to the pixel coordinates in the image.

$$K = \begin{bmatrix} f_u & s & c_u \\ 0 & f_v & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

where f_u, f_v is the focal lengths in column/row direction, s is the skew and c_u, c_v is the center of the image in column/row direction.

In the dataset, each sequence provided a file containing the K matrices for all the sensors in an unordered list. It was therefore assumed that the placement in the list corresponded to the number of the sensor. However, the extrinsic parts of these did not yield correct results. Another set of extrinsic rotation/translation matrices were provided in a different file where the sensor number *was* identified. Still, these extrinsic parameters were calibrated for the color camera of each sensor. Since the depth camera and color camera of the Kinect are placed approximately 20cm from each other, a vector $\begin{bmatrix} 0.02 \\ 0 \\ 0 \end{bmatrix}$ was added after calculating the extrinsic part of the model to compensate. The resulting projections can be seen in figure 3.1c. Still, the K matrices had to be sourced from the unordered list and introduced uncertainty about whether the datasets will be accurate. Since the Kinect cameras are the same model, and after all very similar, this uncertainty were chosen to be ignored.

$$K(x, y) = \begin{cases} \frac{2}{\pi\sigma^2}(1 - ((\frac{x}{\sigma})^2 + (\frac{y}{\sigma})^2)) & \text{if } |(\frac{x}{\sigma})^2 + (\frac{y}{\sigma})^2| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Target limb-maps for each frame were created by projecting all instances of a certain type of limb onto an empty matrix of the same dimensions as the depth frame, illustrated in Figure 3.2. A 2D

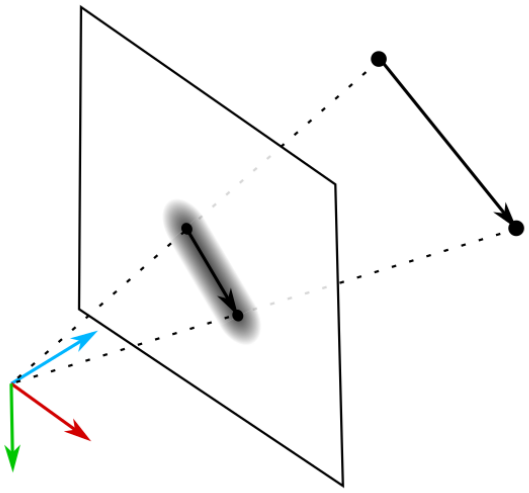
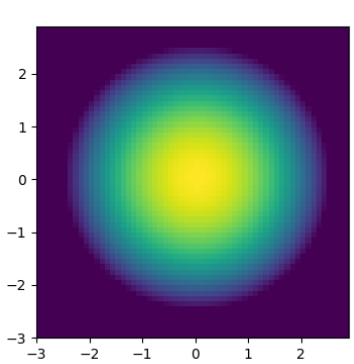
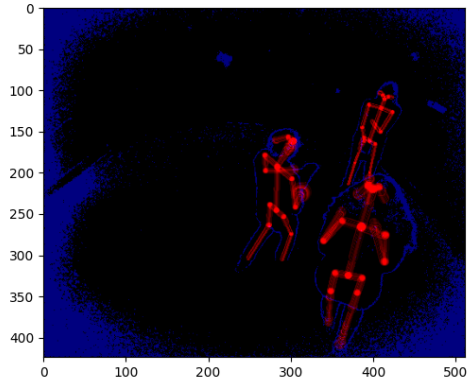


Figure 3.2: A limb is projected onto the image plane



(a) A 2D Epanechnikov [6] kernel.



(b) The magnitudes for the vectors along the limb.

Epanechnikov Kernel with variable σ were then used to calculate the magnitude for each of the 3D vectors pointing in the direction of the limb in question. The limbs furthest away were calculated first, so the vectors from these would be overwritten by subsequent limbs in case they were behind each other. The $\sigma(x, y)$ of the Epanechnikov Kernel was determined by the average depth of a 3×3 kernel around the point. The reason for varying the σ was to simulate that the shell around the limb would appear smaller at a greater distance. The matrix was then thresholded by the magnitude of each vector to generate a cleaner output. However, because the normalization constant would change the magnitude of the vectors at the center according to the distance, the equation was simplified to a quadratic function 3.4.

$$K(x,y) = \begin{cases} 1 - ((\frac{x}{\sigma(x,y)})^2 + (\frac{y}{\sigma(x,y)})^2) & \text{if } |(\frac{x}{\sigma(x,y)})^2 + (\frac{y}{\sigma(x,y)})^2| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

$$\sigma(x,y) = \frac{c}{D(x,y)} \tag{3.5}$$

Equation 3.5 describes how σ changes by the value in the depth image D. The constant c is chosen by the width of the limb in pixels at a certain distance.

The resulting magnitudes for the limbs are illustrated by the declining intensity in figure 3.3b. The produced sample tensors are illustrated in figure 3.4. For each sequence, a "long" tensor is made. The first slice of the tensor is the depth image, of size $1 \times w \times h$. Next, the limb maps are stored. They contain the decomposition of the vector, x, y, z in each pixel. This part of the sample, therefore, has the size $3M \times w \times h$ for M limbs. Last is the Joint Maps, which contain both the distance to the center of the joint and the confidence for each pixel. This results in the size $2N \times w \times h$ for N joints.

A single tensor of samples was created for each sequence. At training time, all sequences except the test sequences were combined to a single dataset which was shuffled.

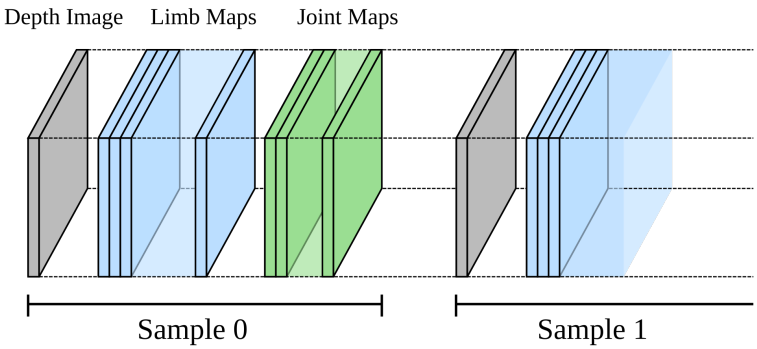


Figure 3.4: The sturcture of the created dataset

Chapter 4

DepthPose

The DepthPose here introduced is a complete system for extracting human 3D poses from a single depth image in real-time[1]. It is specifically tailored for use in robotic applications where computational resources are limited. Because the MECS is envisioned to be a mobile unit, DepthPose has to be robust to be able to see the person from different viewing angles despite possible occlusions between the person and the depth-sensor. The architecture is inspired by and builds on the architecture in OpenPose [3]. The general pipeline of DepthPose follows the OpenPose pipeline in figure 2.2 closely, with the addition of a novel Articulation Network that refines the poses after the bipartite matching.

One of the goals for the MECS project is to look for patterns that could lead to worsening or more dangerous living conditions. To that end, Human Activity Recognition (HAR) is implemented with the purpose of tracking a user from day to day. Representing the pose in 3D will simplify application areas such as HAR, because a 3D representation of a skeleton can be defined by a coordinate system constrained to any two connected limbs from the observed skeletons. This means that two different skeleton observations can be represented by a common reference point. Had the skeletons been represented in 2D, the same skeleton seen from two different angles could look vastly different, even if referenced from the same limb. Comparing the two 2D poses will therefore be a more difficult problem than comparing the same poses referenced from a common 3D coordinate system.

[1] Real-time is defined as being capable of processing at least 30 depth frames per second on the specified minimum hardware requirements.

4.1 Architecture

The system pipeline is outlined in figure 4.1. As in [3], two stages are used to extract the pose from an image. In each stage, a Recurrent Neural Network (RNN) architecture iteratively refines the output from the network at that stage. This iterative architecture was inspired by Convolutional Pose Machines [35].

At timestep $t = 0$ in the first stage, a set of depth features are created by the first CNN, *df*. These are stacked with the *limb-maps* produced by the next CNN, *lp*. At $t = 0$ these limb-maps will be initialized to a known value, $\mathbf{0}$, so only the learned bias weights influence *lp* at this timestep. *lp* will then refine these limb-maps at subsequent timesteps $0 < t \leq T_P$.

At timestep $t = T_P + 1$ the refined "first guess" limb-maps from *lp* are again stacked with the depth features from *df*, and used as inputs to *jp* which produces a set of likely *joint-maps*. These joint-maps are then passed to the Assembly function, which performs the bipartite matching algorithm and constructs a set of skeletons. An instance of the Articulation Network is created for each of the detected skeletons. The Articulation Networks refine the poses for each of the detected skeletons. The refined skeleton poses are then projected onto a set of limb-maps which is used instead of the "first guess" limb-maps from *lp*, in successive timesteps $T_P + 1 < t \leq T_P + T_C$.

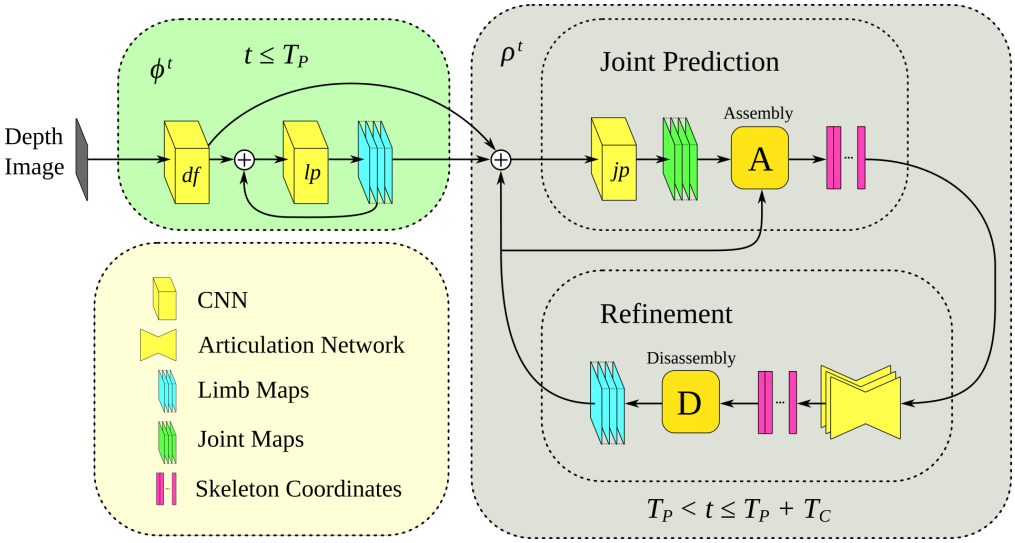


Figure 4.1: Main architecture, as in [4] two recurrent stages, ϕ^t and ρ^t are used to iteratively refine the limb and joint locations.

4.1.1 Depth feature extraction

Figure 4.2: Example of an RGB image on the left vs. a comparative depth image of the same scene on the right. Excerpt sourced from [5].

The main task for the CNN *df* is to find useful depth features for the subsequent networks. As illustrated in Figure 4.2, the information in depth images is not as dense as in RGB images. With the goal of object recognition, networks that are trained on RGB images could rely on features such as colors and textures. For example, the blue shirt vs. the khaki pants or the checkerboard pattern of the floor. However, none of those features are present in the depth image. In both images, edges could be extracted to indicate the outline of objects. Indeed, the edges detected in the depth image could be more useful for this purpose than in the RGB image because they would represent the actual 3D boundary between two objects. A quick example could be two boxes of the same color placed partly behind each other. In an RGB image, they could appear to be part of the same object, whereas a sharp edge would separate them in a depth image. Still, much of the information in depth images have to rely on gradients and other features at a larger scale. It can be speculated that representations for planes geometric or organic shapes will be learned.

Figure 3.1d shows that depth images can also be represented as points in a volume. However, to use such a representation of the data as input to a 3D CNN, a limit would have to be placed on the observed depth to define a fixed input size. This would lead to a sparse representation of the data, where many convolutions would contain empty space. Since no information can be gathered in the occluded parts of the scene anyway, a single channel depth image is a more succinct representation of the input.

As discussed in Section 2.1, the feature extraction parts of a network can be trained on large, unrelated datasets. However, no models that were pre-trained on depth images were found, so applying transfer learning for this part of the network was not an option. This part of the model is therefore not necessarily optimal, and having only seen the inside of a dome, it might not have found features that would perform better in a real-world environment.

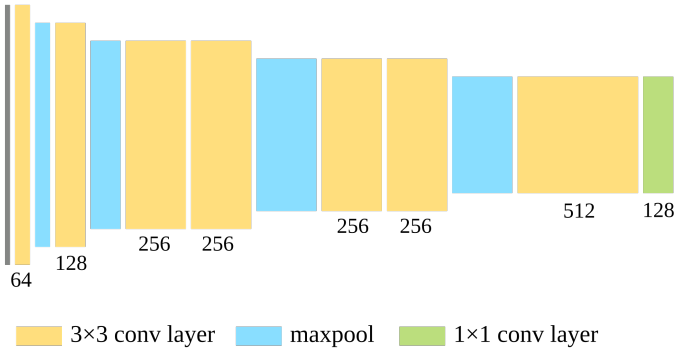


Figure 4.3: Depth feature extractor network. The maxpool layers has size 2×2 and stride 2. Each convolutional layer uses the Rectified Linear Unit (ReLU) activation function.

The OpenPose architecture reuses the first 10 layers of VGG-19 (E configuration) [28]. Therefore, this part of the network takes inspiration from it as well. VGG-19 is constructed using several 3×3 convolutional layers with ReLU [23] activations, and maxpool operations at certain depths. By only using enough 3×3 convolutional layers, the number of learnable parameters are kept down, while preserving the receptive field of a larger kernel (not stacked)[2]. This is ideal for keeping the network as small, and thus as fast as possible. However, since the information at a small scale can be sparse in depth images, dilation is used to obtain a larger receptive field for each convolutional layer. This is different from pooling layers, because they emit which *feature* in a certain layer has the strongest, weakest or calculate what the average response to the input was. On the other hand, a dilated convolutional layer *creates* features for a larger receptive field, and does not downscale the spacial dimensions of the feature tensor, if same-padding is utilized. The first convolutional layer

[2]For an input and output consisting of two channels the learnable parameters can calculated: One 7×7 kernel $\rightarrow 7^2C^2 = 49C^2$ parameters. Three stacked 3×3 kernels $\rightarrow 3(3^2C^2) = 27C^2$ parameters.

in Figure 4.3 uses dilation of 4 to compensate for the lack of small-scale features in the depth image.

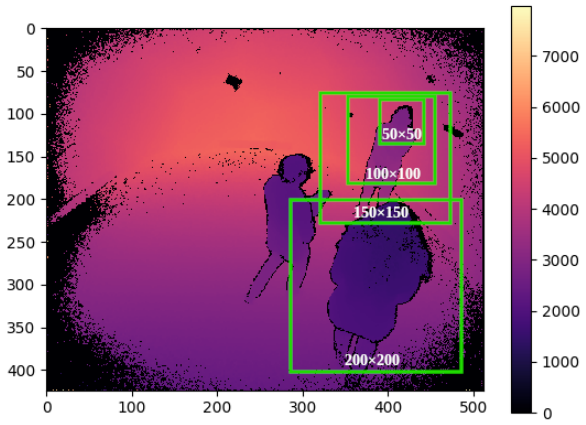


Figure 4.4: Relevant receptive fields in a depth image. Fields are shown as green boxes, with sizes placed near the edge.

4.1.2 Limb- and Joint-Maps

In [3, 35] it is observed that the receptive field of the convolutions are important to establish long-range inferences about body landmarks. This might be because the network learns about the natural symmetries in the human body, and can therefore more easily predict the relationship between different body parts.

In the revisited version of OpenPose, the 7×7 convolution blocks in the pipeline are replaced by convolution blocks of three 3×3 filters, where the emissions from each filter is concatenated at the end of the block. This preserves the receptive field of the convolution block, while reducing the number of learnable parameters, as noted in [28]. In addition, by concatenating the result from each layer at the end of the block, a type of residual network [12] is created, mitigating the vanishing gradient problem.

DepthPose uses three such convolution blocks, with a two final 1×1 convolution layers with dropout between them. This final feature vector is then upsampled through a transposed convolution layer. The output of this is a tensor of $w \times h \times N$ where N is the number of feature maps needed for the network. Since the limb and joint maps look for the same kind of

long-reaching features, the architecture is similar.

The limb-map network produce tensors $3M \times w \times h$ where M is the number of limbs in a skeleton. Each limb has 3 components, as they encode the vector pointing in the direction of the limb at that coordinate.

The joint-maps are tensors $2N \times w \times h$ where N is the number of joints in a skeleton. The tensors encode the depth of the limb at that location, as well as the confidence.

4.1.3 Articulation network

The articulation network solves the problem where two people occludes the other. If a joint of the same body part for two people is co-linear to the camera, this joint can not be represented in the joint-map for the occluded person.

The articulation network is stacked on top of the part-detection network and its main role is to refine the limb lengths and angles between each joint. Each of the detected persons are passed through the articulation network, which leads to a bit more complexity and runtime for the network based on the number of people. However, since the network has so comparatively few inputs, and is quite shallow, preformance is not expected to suffer notably.

The coordinates and confidences for each joint (if not detected, confidence is 0) is the input to the network. The network will try to find out what the positions of joints with low confidences, or no detections, should be. It is hypothesized that this network will learn things like symmetry (left and right limbs should have the same length), proportionality (limbs should be proportional to each other), possible articulations, and natural poses.

The architecture is visualized as a simple, almost fully connected neural network. Since each joint has four properties, (x, y, z, c), these are input to a single neuron in the network. The layers after this is however fully connected.

4.2 Assembly

Done in the same way as [3], however the score-equation for the line integral is

$$S = \sum_{i=0}^{K} \cos(\theta) \tag{4.1}$$

θ is the angle between the the vector defined by the candidate points and the angle of the vector at the points of the limb-maps along the line i, K

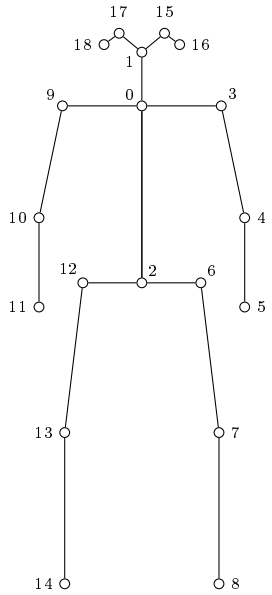


Figure 4.5: Numbering for detected landmarks/keypoint markers.

ID	Description	Std.Coord.
0	Neck	(0.00, 2.34)
1	Nose	(0.00, 3.05)
2	Middle hip	(0.00, 0.00)
3	Left shoulder	(1.05, 2.34)
4	Left elbow	(1.36, 0.86)
5	Left wrist	(1.36, -0.32)
6	Left hip	(0.78, 0.00)
7	Left knee	(1.02, -1.98)
8	Left ankle	(1.02, -3.98)
9	Right shoulder	(-1.05, 2.34)
10	Right elbow	(-1.36, 0.86)
11	Right wrist	(-1.36, -0.32)
12	Right hip	(-0.78, 0.00)
13	Right knee	(-1.02, -1.98)
14	Right ankle	(-1.02, -3.98)
15	Left eye	(0.30, 3.30)
16	Left ear	(0.50, 3.15)
17	Right eye	(-0.30, 3.30)
18	Right ear	(-0.50, 3.15)

Table 4.1: Numberings, names/descriptions and standard coordinates for recognized landmarks

4.3 Training

Chapter 5

Experiments

- compare the learnable parameters between the openpose architecture and the depthpose architecture.
- calculate the euler distance between joint-locations output by the pipeline and the ground-truth locations (skeleton matching must be preformed, since it is not guaranteed that the IDs for the skeletons would match)
- compare accuracy and complexity with different values for T_P and T_C
- assess the usefullness of the articulation network by comparing the joint locations after refinement with joint locations output at $t = T_P + 1$, where just limb-maps from the first stage are used in the prediction.
- could T_P be reduced to achieve usefulness of the articulation network?

5.1 Results

Chapter 6

Conclusions

Currently no conclusions can be drawn from the lack of experiments. However, It should be possible to use the articulation network to provide complete poses, even if key landmarks are occluded.

Chapter 7

Future Work

- Find a more accurate pose from a temporal algorithm that takes input from a series of depth images from a single viewpoint.
- Combine multiple viewpoints for a more accurate pose.
- Human activity recognition using 3D pose provided by the method proposed in this paper.
- Train the network over a larger dataset in unstructured environments and with multiple people present.

Appendices

A Depth sensors

Since depth sensors are widely used in different robotics applications for tasks such as SLAM, odometry and object detection, we selected this as our main source of information for monitoring the user. There are mainly three different technologies to choose from: *Structured light*, *Time-of-Flight (ToF)* and *Stereo vision*.

A.1 Stereo vision

uses two cameras that are observing parts of the same scene. In commercial packages the cameras are usually calibrated, so we have measurements to put into the camera matrix as well as the rotation and translation between the two camera matrices.

However, to get a 3D structure, we need to find common feature points between the two cameras. To do this, we can use various feature descriptors such as ORB, SWIFT and SURF. When good matches has been found between the images, we measure the disparity between the points, and triangulate the distance. The depth measurements for the rest of the image are then calculated by matching pixels close to the found featurepoints.

Since this is an optically based technology, it will work well in well-lit scenes that contain many unique featurepoints. If we operate in an homogenous environment with few, or similar textures it will be difficult to find featurepoints to map the environment. An example of this could be on the seabed or inside buildings with limited light conditions, for example during a blackout.

A.2 Structured light

uses a projected pattern of light points onto the scene which is registered by a calibrated camera. Usually, the projected light pattern and camera operate in the infrared part of the electromagnetic spectrum[1]. This means that in locations where one can expect a lot of IR radiation, this technology will not work very well. Since the IR radiation from the sun usually is much stronger than the one emitted from the projector on the sensor, this technology will not work well outside in well-lit conditions. It will however work inside and in conditions where no external light source are provided.

[1] The Microsoft Kinect V2 sensor uses a wavelength of 827-850nm, according to [10, Chapter 4.1]

In addition, since the light is structured and the sensor is calibrated, we can skip the step where we find common featurepoints to triangulate the distance which we have to do in the stereo vision case.

A.3 Time-of-Flight

cameras uses the known constant of c to calculate distances in the image, by measuring the time a light-pulse emitted from the camera uses to be reflected onto the camera sensor. For example, Microsofts Kinect v2 uses a specialized ToF-pixel array in conjunction with a timing generator and modulated laser diodes to obtain per-pixel depth images [26].

As with structured light sensors, this is susceptible to interferrence from external light sources, or specular surfaces, and has limited range because of light fall-off. However, since the distance calculations are timing based, we can obtain framerates up to 30 fps in the Kinect v2 sensor [19].

B Robotic Operating System

In order to make the system easier to use and available to as many platforms as possible, it was decided to create it for the Robotic Operating System (ROS). ROS is a collection of libraries and a runtime environment making communication with different modules and programs on the robot possible.

C Classifiers

write a bit about what classifiers are, and how we use them to find the different keypoints in the image.

Bibliography

[1] Mykhaylo Andriluka et al. '2D Human Pose Estimation: New Benchmark and State of the Art Analysis'. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014.

[2] H. Bay, T. Tuytelaars and L. Van Gool. 'SURF: Speeded Up Robust Features'. In: *Computer Vision – ECCV 2006*. Ed. by A. Leonardis, H. Bischof and A. Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417. ISBN: 978-3-540-33833-8.

[3] Zhe Cao et al. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*. 2019. arXiv: 1812.08008 [cs.CV].

[4] Zhe Cao et al. 'Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields'. In: *CVPR*. 2017.

[5] Maxime Devanne. '3D Human Behavior Understanding by Shape Analysis of Human Motion and Pose'. PhD thesis. Dec. 2015.

[6] V. A. Epanechnikov. 'Non-Parametric Estimation of a Multivariate Probability Density'. In: *Theory of Probability & Its Applications* 14.1 (1969), pp. 153–158. DOI: 10.1137/1114019. eprint: https://doi.org/10.1137/1114019. URL: https://doi.org/10.1137/1114019.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. 'Pictorial Structures for Object Recognition'. In: *Int. J. Comput. Vision* 61.1 (Jan. 2005), pp. 55–79. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000042934.15159.49. URL: https://doi.org/10.1023/B:VISI.0000042934.15159.49.

[8] M.A. Fischler and R.A. Elschlager. 'The Representation and Matching of Pictorial Structures'. In: *IEEE Transactions on Computers* C-22.1 (1973), pp. 67–92. DOI: 10.1109/T-C.1973.223602.

[9] Kunihiko Fukushima. 'Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position'. In: *Biological Cybernetics* 36.4 (Apr. 1980), pp. 193–202. ISSN: 1432-0770. DOI: 10.1007/BF00344251. URL: https://doi.org/10.1007/BF00344251.

[10] Silvio Giancola, Matteo Valenti and Remo Sala. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springer International Publishing, 2018. DOI: 10.1007/978-3-319-91761-0. URL: http://dx.doi.org/10.1007/978-3-319-91761-0.

[11] Grauman, Shakhnarovich and Darrell. 'Inferring 3D structure with a statistical image-based shape model'. In: *Proceedings Ninth IEEE International Conference on Computer Vision*. 2003, 641–647 vol.1. DOI: 10.1109/ICCV.2003.1238408.

[12] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[13] P. V. C. Hough. 'Method and means for recognizing complex patterns'. US 3 069 654A. 1960. URL: https://patents.google.com/patent/US3069654A/en.

[14] Catalin Ionescu et al. 'Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1325–1339.

[15] Hanbyul Joo et al. 'Panoptic Studio: A Massively Multiview System for Social Interaction Capture'. In: 2017.

[16] Hanbyul Joo et al. 'Panoptic Studio: A Massively Multiview System for Social Motion Capture'. In: *ICCV*. 2015.

[17] Helseetaten Oslo Kommune. *Oslohelsa – Kortversjonen, Oversikt over helsetilstand og påvirkningsfaktorer*. June 2016. URL: https://www.oslo.kommune.no/getfile.php/13139280/Innhold/Politikk%20og%20administrasjon/Statistikk/Oslohelsa_kortversjon.pdf.

[18] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[19] Elise Lachat et al. 'Assessment and Calibration of a RGB-D Camera (Kinect v2 Sensor) Towards a Potential Use for Close-Range 3D Modeling'. In: *Remote Sensing* 7.10 (Oct. 2015), pp. 13070–13097. ISSN: 2072-4292. DOI: 10.3390/rs71013070. URL: http://dx.doi.org/10.3390/rs71013070.

[20] D. G. Lowe. 'Object recognition from local scale-invariant features'. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2. 1999, 1150–1157 vol.2.

[21] N. Magnenat-Thalmann, R. Laperrière and D. Thalmann. 'Joint-Dependent Local Deformations for Hand Animation and Object Grasping'. In: *Proceedings on Graphics Interface '88*. Edmonton, Alberta, Canada: Canadian Information Processing Society, 1989, pp. 26–33.

[22] Greg Mori and Jitendra Malik. 'Estimating Human Body Configurations Using Shape Context Matching'. In: *Proceedings of the 7th European Conference on Computer Vision-Part III*. ECCV '02. Berlin, Heidelberg: Springer-Verlag, 2002, pp. 666–680. ISBN: 3540437460.

[23] Vinod Nair and Geoffrey E. Hinton. 'Rectified Linear Units Improve Restricted Boltzmann Machines'. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 9781605589077.

[24] D. Ramanan and D.A. Forsyth. 'Finding and tracking people from the bottom up'. In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Vol. 2. 2003, pp. II–II. DOI: 10.1109/CVPR.2003.1211504.

[25] E. Rublee et al. 'ORB: An efficient alternative to SIFT or SURF'. In: *2011 International Conference on Computer Vision*. 2011, pp. 2564–2571.

[26] John Sell and Pat O'Connor. 'XBOX One Silicon'. Hot Chips 25. Aug. 2013. URL: http://www.hotchips.org/wp-content/uploads/hc_archives/hc25/HC25.10-SoC1-epub/HC25.26.121-fixed-%20XB1%2020130826gnn.pdf.

[27] Jamie Shotton et al. 'Efficient Human Pose Estimation from Single Depth Images'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2013), pp. 2821–2840. DOI: 10.1109/TPAMI.2012.241.

[28] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[29] Xinhang Song, Luis Herranz and Shuqiang Jiang. *Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs*. 2018. arXiv: 1801.06797 [cs.CV].

[30] O. Tange. 'GNU Parallel - The Command-Line Power Tool'. In: *;login: The USENIX Magazine* 36.1 (Feb. 2011), pp. 42–47. DOI: http://dx.doi.org/10.5281/zenodo.16303. URL: http://www.gnu.org/s/parallel.

[31] Jonathan Tompson et al. *Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation*. 2014. arXiv: 1406.2984 [cs.CV].

[32] S. Ullman and S. Brenner. 'The interpretation of structure from motion'. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 203.1153 (1979), pp. 405–426. DOI: 10.1098/rspb.1979.0006. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.1979.0006. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1979.0006.

[33] P. Viola and M. Jones. 'Rapid object detection using a boosted cascade of simple features'. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. 2001, pp. I–I.

[34] Keze Wang et al. 'Human Pose Estimation from Depth Images via Inference Embedded Multi-Task Learning'. In: *Proceedings of the 24th ACM International Conference on Multimedia*. MM '16. Amsterdam, The Netherlands: Association for Computing Machinery, 2016, pp. 1227–1236. ISBN: 9781450336031. DOI: 10.1145/2964284.2964322. URL: https://doi.org/10.1145/2964284.2964322.

[35] Shih-En Wei et al. 'Convolutional pose machines'. In: *CVPR*. 2016.