Mendel University in Brno
Faculty of Business and Economics

Automated neural network learning for higher accuracy human skeleton detection under realistic conditions

Master's Thesis

Bc. Damián Sova

Supervisor:
doc. Ing. Oldřich Trenz, Ph.D.

Brno 2024

MENDELU
Faculty
of Business
and Economics

Department of Informatics
Academic year: 2023/2024

DIPLOMA THESIS TOPIC

Author of thesis:	**Damián Sova**
Study programme:	Open Informatics
Scope:	Scope for Open Informatics
Topic:	**Automated neural network learning for higher accuracy human skeleton detection under realistic conditions**
Length of thesis:	2,5–4 AA

Guides to writing a thesis:

1. The aim of this thesis is to analyze the problem of image detection using neural networks and to design a neural network model for human skeleton detection in real conditions.
2. Analyze the current state of the art in human skeleton detection, i.e., methods, approaches, available datasets.
3. Design a neural network model using available approaches (BodyPoseNet – NVIDIA, MediaPipe PoseNet – Google, ViTPose) for human skeleton detection. As part of the solution to this item, create a reference dataset (image dataset for skeleton detection).
4. Implement the proposed solution using freely available technologies. Use the Apple iPhone 14 Pro, Lidar, as a possible extension in the 3D level.
5. Evaluate your own solution and formulate options for further development.

Selected bibliography:

1. ARLOW, Jim; NEUSTADT, Ila. *UML 2 a unifikovaný proces vývoje aplikací: objektově orientovaná analýza a návrh prakticky.* 2nd ed. Brno: Computer Press, 2007. 567 p. ISBN 978-80-251-1503-9.
2. PATTON, Ron. *Software Testing.* Indiana: Sams Publishing, 2005. 408 p. ISBN 978-0-672-32798-8.
3. CHOLLET, François; PECINOVSKÝ, Rudolf. *Deep learning v jazyku Python: knihovny Keras, Tensorflow.* 1st ed. Praha: Grada Publishing, 2019. 328 p. Knihovna programátora. ISBN 978-80-247-3100-1.
4. CHOLLET, François. *Deep learning with Python.* Shelter Island: Manning, 2021. 478 p. ISBN 978-1-61729-686-4.
5. C. Patil and V. Gupta (2021, July 15). Human pose estimation using keypoint RCNN in pytorch. LearnOpenCV. https://learnopencv.com/human-pose-estimation-using-keypoint-rcnnin-pytorch/.
6. Rosebrock, A. (2021, April 17). R-CNN object detection with Keras, tensorflow, and Deep Learning. PyImageSearch. https://pyimagesearch.com/2020/07/13/r-cnn-object-detection-withkeras- tensorflow-and-deep-learning/.
7. Z. Tang, D. Wang and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5900-5904, doi: 10.1109/ICASSP.2016.7472809.

Diploma thesis topic submission date: May 2023

Deadline for submission of Diploma thesis: May 2024

L. S.

Electronic approval: 29. 6. 2023
doc. Ing. Oldřich Trenz, Ph.D.
Thesis supervisor

Electronic approval: 29. 6. 2023
Damián Sova
Author of thesis

Electronic approval: 29. 6. 2023
prof. Ing. Cyril Klimeš, CSc.
Head of Institute

Electronic approval: 29. 6. 2023
doc. Ing. František Dařena, Ph.D.
Study programme supervisor

I express my sincere gratitude to my supervisor, Doc. Ing. Oldřich Trenz, Ph.D., for his invaluable time, unwavering support, and insightful guidance throughout the entirety of this research journey. I am also deeply appreciative of the expert consultations provided by RNDr. Michal Procházka, Ph.D., from visioncraft s.r.o., whose regular insights played a pivotal role in shaping the trajectory of this work. Lastly, heartfelt thanks to my friends and family for their unwavering support, encouragement, and understanding, without which this endeavor would not have been possible.

Declaration

I hereby declare that this thesis entitled *Automated neural network learning for higher accuracy human skeleton detection under realistic conditions* was written and completed by me. I also declare that all the sources and information used to complete the thesis are included in the list of references. I agree that the thesis could be made public in accordance with Article 47b of Act No. 111/1998 Coll., Higher Education Institutions and on Amendments and Supplements to Some Other Acts (the Higher Education Act), and in accordance with the current Directive on publishing of the final thesis. I declare that the printed version of the thesis and electronic version of the thesis published in the application Final Thesis in the University Information System is identical.

I am aware that my thesis is written in accordance to Act No. 121/2000 Coll. (the Copyright Act) and therefore Mendel University in Brno has the right to conclude licence agreements on the utilization of the thesis as a school work in accordance with Article 60 (1) of the Copyright Act.

Before concluding a licence agreement on utilization of the work by another person, I will request a written statement from the university that the licence agreement is not in contradiction to legitimate interests of the university, and I will also pay a prospective fee to cover the cost incurred in creating the work to the full amount of such costs.

Brno January 29, 2024 ....

signature

Abstract

Sova, Damián. *Automated neural network learning for higher accuracy human skeleton detection under realistic conditions.* Master's Thesis. Brno : Mendel University in Brno, 2024.

Key words

image processing, pose estimation, human skeleton detection

Abstrakt

Sova, Damián. *Automatizované učenie neurónových sietí na presnejšiu detekciu ľudskej kostry v reálnych podmienkach.* Diplomová práca. Brno : Mendelova univerzita v Brně, 2024.

Kľúčové slová

spracovanie obrazu, odhadovanie pózy, detekcia ľudskej kostry

Contents

1 Introduction **9**

1.1 Motivation and Basic Objectives of the Work 9
1.2 Current State and Problem to Be Addressed 10

2 Theoretical foundations **11**

2.1 Neural Network 11
2.1.1 How Neural Network Works 11
2.2 Convolutional Neural Network 12
2.2.1 How Convolutional Layers Work 13
2.2.2 Pooling Layers 14
2.2.3 Fully Connected Layers 14
2.2.4 Training the CNN 14
2.2.5 Example of CNN Usage 15
2.2.6 Limitations of Current Methods 15
2.3 Region-based Convolutional Neural Network 15
2.4 Existing s for Human Pose Estimation 16
2.5 Posenet 17
2.6 Movenet 18
2.7 MMPose 19
2.8 Chapter Summary 20

3 Practical part **21**

3.1 Dataset 21
3.2 Created Unified Format 21
3.3 Experiments and Results 21
3.4 Implementation Problems and Technical Limitations 22

4 Conclusion **23**

References **24**

List of Tables **26**

List of Figures 27

List of Abbreviations 28

List of Source Codes 29

APPENDICES

1 Introduction

1.1 Motivation and Basic Objectives of the Work

The field of *computer vision* has witnessed rapid evolution, serving as the foundation for understanding visual information in images and videos (Szeliski, 2010). Within this context, the accurate *detection* of the *human skeleton* holds immense potential for applications ranging from autonomous systems to healthcare. The motivation driving this master's thesis is to *enhance* the *precision* of human skeleton detection under *realistic* conditions through the application of *automated* neural network learning.

The primary objectives of this work can be delineated as follows:

- **Technical Challenges in Neural Network Training:** Training a neural network (NN) for human skeleton detection is inherently challenging. It necessitates the availability of hardware capable of capturing the human body's spatial position through sensors placed on key body points, which are crucial for the detection process. Acquiring sensor data is essential for constructing a comprehensive training dataset for the NN. However, the generation of training data often occurs in controlled "laboratory conditions," using props and actors (Yang, 2018). Consequently, the creation of such a model becomes resource-intensive, requiring significant investments in time, computational resources, human effort, and hardware. Furthermore, the model's accuracy is constrained by the level of correlation between simulated activities and real-world conditions in the detected scenario.
- **Refinement of Neural Networks for Skeleton Detection:** Existing models for human skeleton detection exhibit limited accuracy for specific use cases due to training in artificially created conditions (Toshev et al., 2014). The proposed approach involves leveraging existing NN models and combining their functionalities without intervention or retraining. To construct a training dataset, real-world data, such as videos capturing falls in nursing homes, will be used. Existing NN models will extract information about the skeleton from these data, which will then be utilized to train a new model with the aim of enhancing accuracy.
- (Optional) **Dimensional Enhancement for Improved Detection:** In scenarios where body position is not clearly visible, particularly when extracting skeleton data from videos without body position sensor data,

inaccurate detection may occur. To address this, the training dataset will be expanded into a three-dimensional space using a lidar sensor on the iPhone 14 PRO. The addition of a third dimension aims to refine skeleton detection in situations where only two-dimensional data are available.

1.2 Current State and Problem to Be Addressed

At present, there is a notable gap in tools and methodologies dedicated to training models for human skeleton detection, utilizing pre-existing models (Yang et al., 2016). While various tools exist for model optimization, compression, and transfer learning to different models, there is a lack of knowledge regarding approaches that integrate existing NNs for training entirely new models. This thesis aims to bridge this gap by exploring the combination of existing neural networks to train a novel model specifically for human skeleton detection, addressing the current limitations in accuracy and practicality associated with conventional training methodologies.

2 Theoretical foundations

This chapter provides an overview of the theoretical foundations of the proposed automated NNs learning approach for human skeleton detection. It introduces the key concepts of NNs, convolutional neural network (CNN), region-based convolutional neural network (RCNN), and transformation models of NN-s. Additionally, it explores existing NNs for human pose estimation, including *Posenet*, *Movenet*, and *MMPose*.

2.1 Neural Network

NNs, inspired by the structure and function of the *human brain*, are computational models comprising *interconnected* layers of artificial *neurons* responsible for processing and transforming information. Demonstrating remarkable capabilities, NNs have proven effective in diverse tasks, including image recognition, natural language processing, and machine translation. A schematic representation of a simple NN is presented in **Figure 2.1**, illustrating individual layers of neurons interconnected with their neighbors. The initial layer is commonly referred to as the *input layer*, followed by *hidden layers*, and concluding with the *output layer*. In practical usage, data, such as an image in the form of a vector where values represent individual pixels, is input into the initial layer for analysis. The NN processes this information, ultimately yielding a result in the form of a single value or vector, dependent on the nature of the problem—be it a classification or regression task. Across various fields, NNs have consistently demonstrated their robustness, excelling in tasks such as classification, prediction, filtering, optimization, pattern recognition, and function approximation (Simoneau et al., 1998).

2.1.1 How Neural Network Works

A NN inspired by the human brain, is a computational system organized into layers of artificial neurons (Nielsen, 2015). Each connection between neurons has a *weight*, representing the strength of influence (Goodfellow et al., 2016). The network learns by adjusting these weights during training, where it processes input data through layers, utilizes *activation functions* to determine

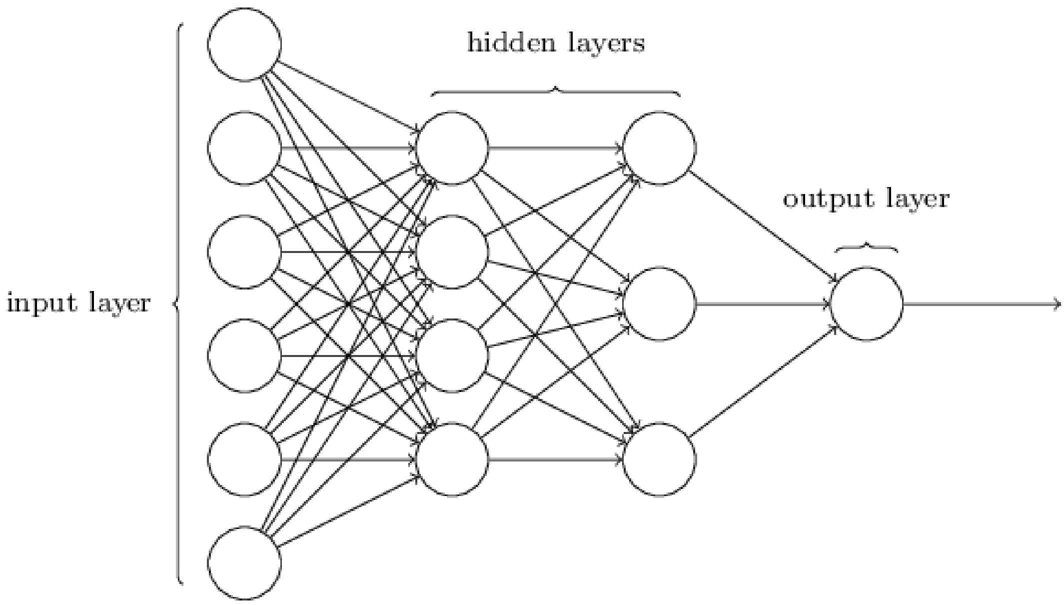

Figure 2.1
Example neural network schema. Source: (Nielsen, 2015)

neuron 'firing', and iteratively adjusts weights based on the difference between predicted and actual outcomes (Nielsen, 2015; Goodfellow et al., 2016; Mazur, 2015). The forward pass involves making predictions, while the backward pass compares predictions to actual results, adjusting weights to minimize *errors* (Mazur, 2015). This learning process enables the neural network to recognize patterns and make accurate decisions in tasks like *image recognition* or *language processing* (Goodfellow et al., 2016).

2.2 Convolutional Neural Network

CNNs are a type of NN architecture that excels at processing and analyzing visual data, such as images and videos. They are particularly well-suited for skeleton detection due to their ability to *extract* local features from the input data. CNNs typically consist of a series of *convolutional layers*, each of which applies a *filter* or *kernel* to the input data to extract *features*. The filters are learned during the training process, allowing the CNN to learn the patterns and relationships that are important for skeleton detection (Singh, 2019). For better understanding of the CNN architecture see example **Figure 2.2**.

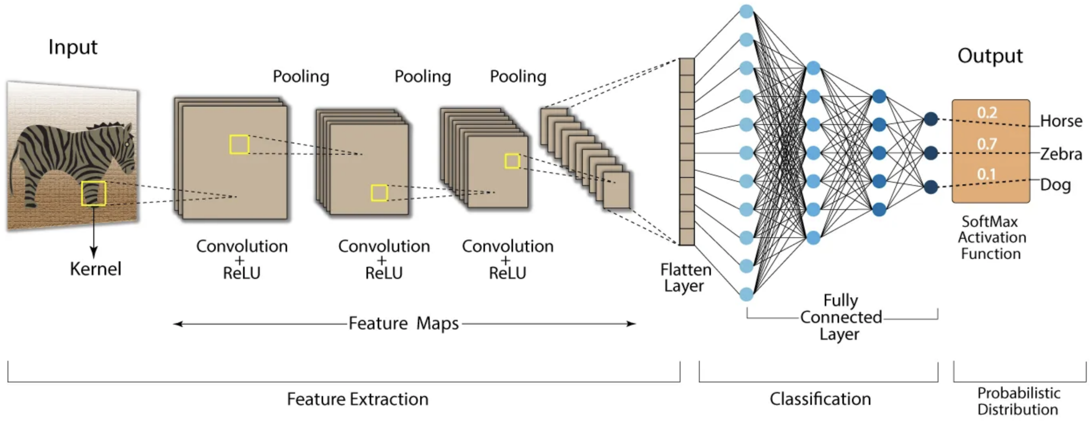

Figure 2.2
A simple classification architecture by CNN. Source: (Koushik, 2023)

CNNs have several advantages for skeleton detection (Huang, 2022):

- **Translation Invariance:** CNNs are invariant to small translations in the input data. This is important for skeleton detection, as the human body can be in *different positions* in an image or video.
- **Feature Learning:** CNNs can learn *complex features* from the input data, which is essential for accurate skeleton detection.
- **Parameter Sharing:** CNNs share *weights* across different positions in the input data. This reduces the number of parameters in the network, making it more efficient and easier to train.

CNNs have become the dominant architecture for skeleton detection, and they have significantly improved the accuracy of this task (Singh, 2019 Huang, 2022).

2.2.1 How Convolutional Layers Work

Each convolutional layer in a CNN takes an input image and applies a filter to it to extract features. The filter is a small matrix of weights that slides across the input image, producing a feature map at each position. The feature map is a representation of the input image that highlights the patterns that are relevant to the task at hand (Agarwal et al., 2019).

For example, in the case of human skeleton detection, a filter might be used to extract features that are indicative of human joints, such as the elbows, knees, and wrists. The feature map produced by this filter would highlight the locations of these joints in the input image.

2.2.2 Pooling Layers

After the convolutional layers extract features, pooling layers are often used to reduce the dimensionality of the feature maps. This helps to reduce the computational cost of the network and also helps to make the network more invariant to small changes in the input data.

Pooling layers work by dividing the feature map into smaller regions and then taking the maximum or average value of each region. This produces a smaller feature map that still contains the most important features from the original image (Agarwal et al., 2019).

2.2.3 Fully Connected Layers

Once the feature maps have been extracted and pooled, they are passed through a series of fully connected layers. These layers are similar to the artificial neurons that are found in traditional neural networks. They take an input vector and produce an output vector.

In the case of human skeleton detection, the fully connected layers are used to classify the detected features as either human joints or background. The output vector from the final fully connected layer is a probability distribution over the possible classes (Agarwal et al., 2019).

2.2.4 Training the CNN

The CNN is trained using a process called supervised learning. This involves providing the network with a dataset of labeled images, where each image is labeled with the positions of the human joints. The network then learns to associate the features extracted from the images with the corresponding labels.

The training process involves adjusting the weights of the filters and connections in the network. This is done using an algorithm called backpropagation, which iteratively updates the weights to minimize the error between the network's predictions and the ground truth labels (Agarwal et al., 2019).

2.2.5 Example of CNN Usage

To illustrate how a CNN is used for human skeleton detection, consider a scenario where a CNN is tasked with detecting human skeletons in a video stream. The CNN would first extract features from each frame of the video using its convolutional layers. Then, it would use these features to predict the positions of the human joints in the frame.

2.2.6 Limitations of Current Methods

While CNNs have achieved significant success in human skeleton detection, there are still some limitations to these methods. One limitation is that CNNs can be computationally expensive, especially when dealing with high-resolution images or videos. Additionally, CNNs can be sensitive to noise and occlusions, which can make it difficult to accurately detect skeletons in real-world scenarios.

Researchers are continuing to develop new methods to improve the accuracy and efficiency of CNNs for human skeleton detection. These methods include using deeper networks, exploring new architectures, and developing more efficient training algorithms (Agarwal et al., 2019).

2.3 Region-based Convolutional Neural Network

RCNNs are a class of deep CNNs that have been widely used for object detection and localization. They are typically characterized by a two-stage pipeline that involves region proposal and region classification (Ren et al., 2015). In the **Figure 2.3** is displayed possible detection scenario of the RCNN.

- **Region Proposal:** The first stage of an RCNN involves generating a set of region proposals, which are candidate bounding boxes for objects in the input image. These proposals are typically generated using a selective search algorithm (He et al., 2015) that identifies regions that are likely to contain objects based on their visual saliency and spatial context (Girshick et al., 2016).

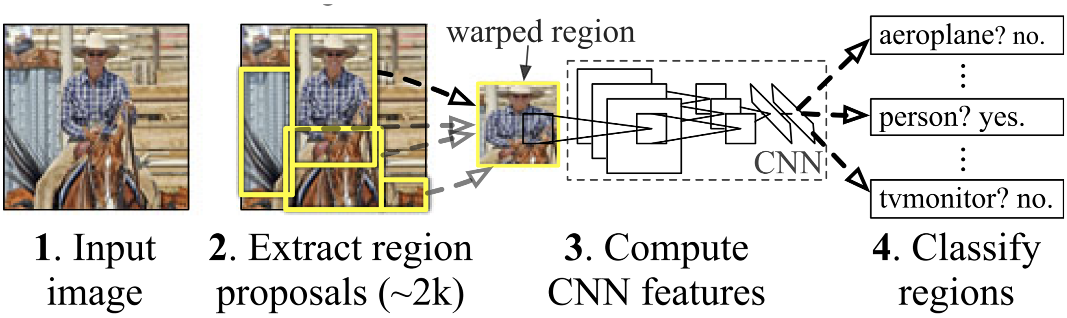

Figure 2.3
RCNN stages. Source: (Girshick, 2016)

- **Feature Extraction and Classification:** The second stage of an RCNN involves classifying each region proposal as either containing the object or not (Ren et al., 2015). This is accomplished by using a CNN to extract feature vectors from each proposal and then applying a classifier to determine whether the features are indicative of the object (Girshick et al., 2016).

The original RCNN architecture has been criticized for its computational inefficiency, as it involves two separate stages of processing (Ren et al., 2015). To address this issue, researchers developed Faster R-CNN, which integrates the region proposal and region classification stages into a single network (Ren et al., 2015). This significantly reduces the computational cost and improves the overall performance of the system (He et al., 2015).

2.4 Existing NNs for Human Pose Estimation

Several NN architectures have been developed for skeleton detection. Here are three notable examples:

(1) **Posenet** (Mediapipe): A lightweight and efficient NN for human pose estimation. It uses a single-stage architecture and can run on mobile devices.
(2) **Movenet** (TensorFlow): A multimodal NN that combines pose estimation, hand tracking, and object tracking. It offers a variety of models with different tradeoffs between accuracy and speed.
(3) **MMPose** (Open-MMLab): A modular and extensible library for pose estimation. It provides a wide range of models and training tools.

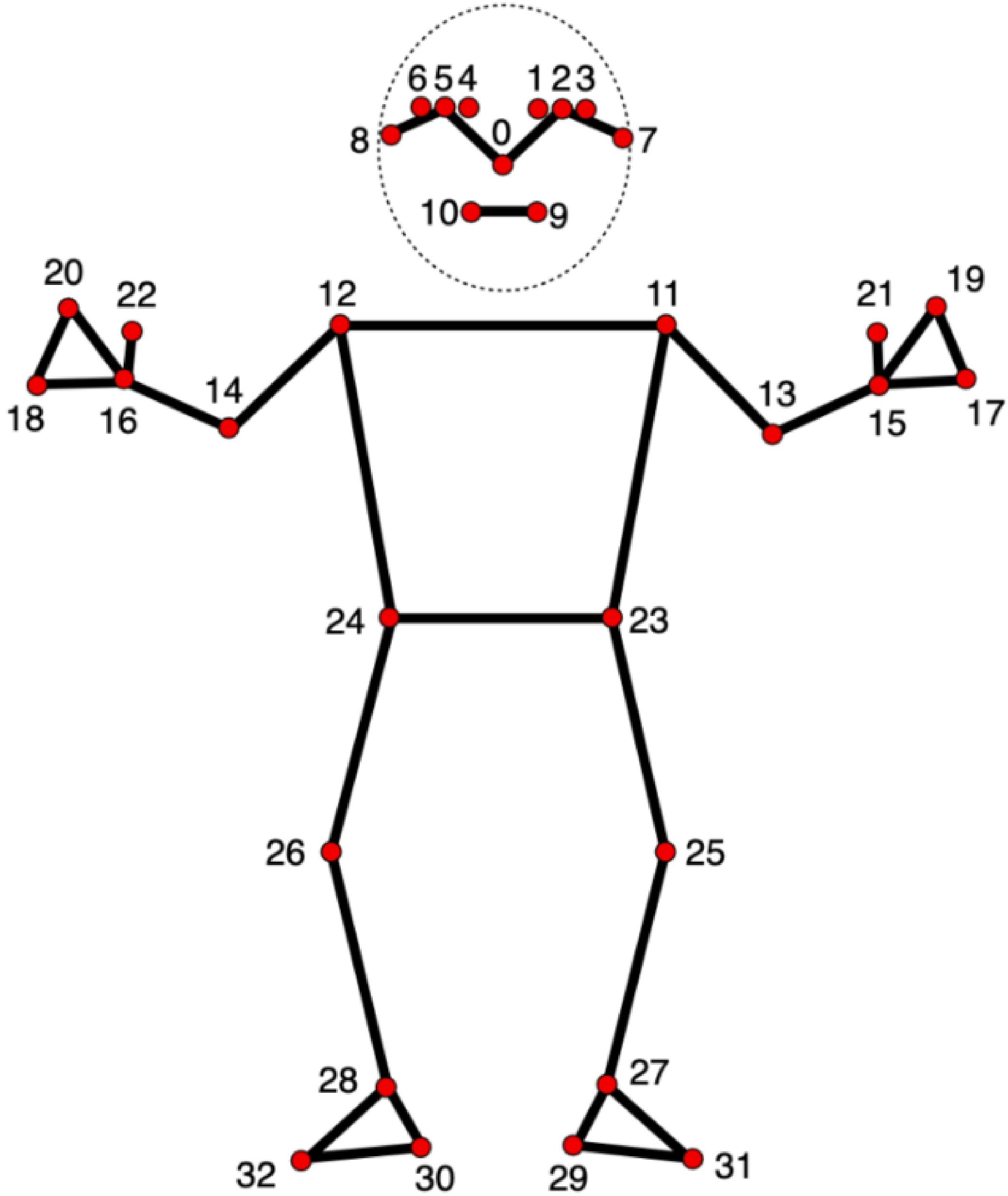

Figure 2.4
Posenet skeleton structure with IDs to each keypoint

2.5 Posenet

The ouput structure of the *Posenet* model can be found in **Figure 2.4**.

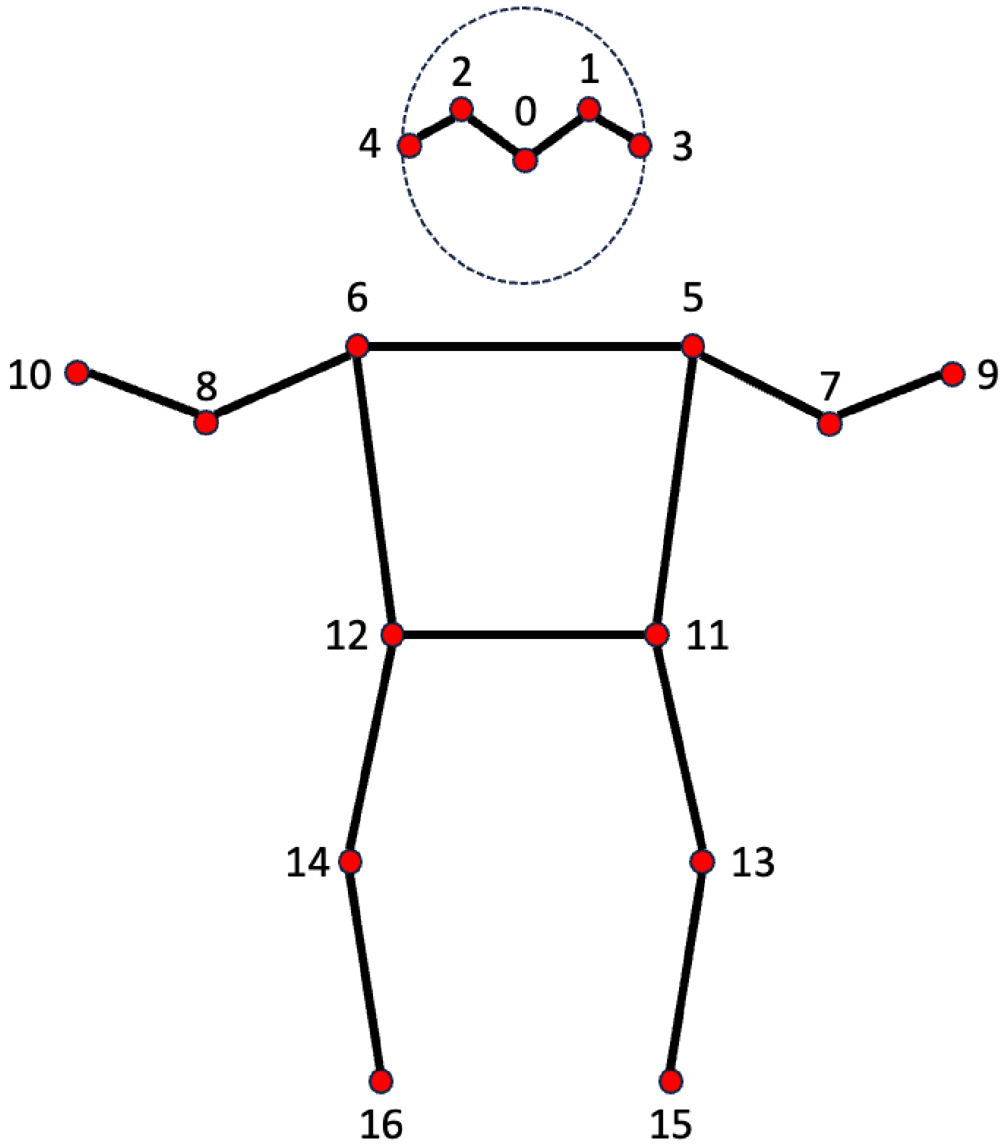

Figure 2.5
Movenet skeleton structure with IDs to each keypoint

2.6 Movenet

The ouput structure of the *Movenet* model can be found in **Figure 2.5**.

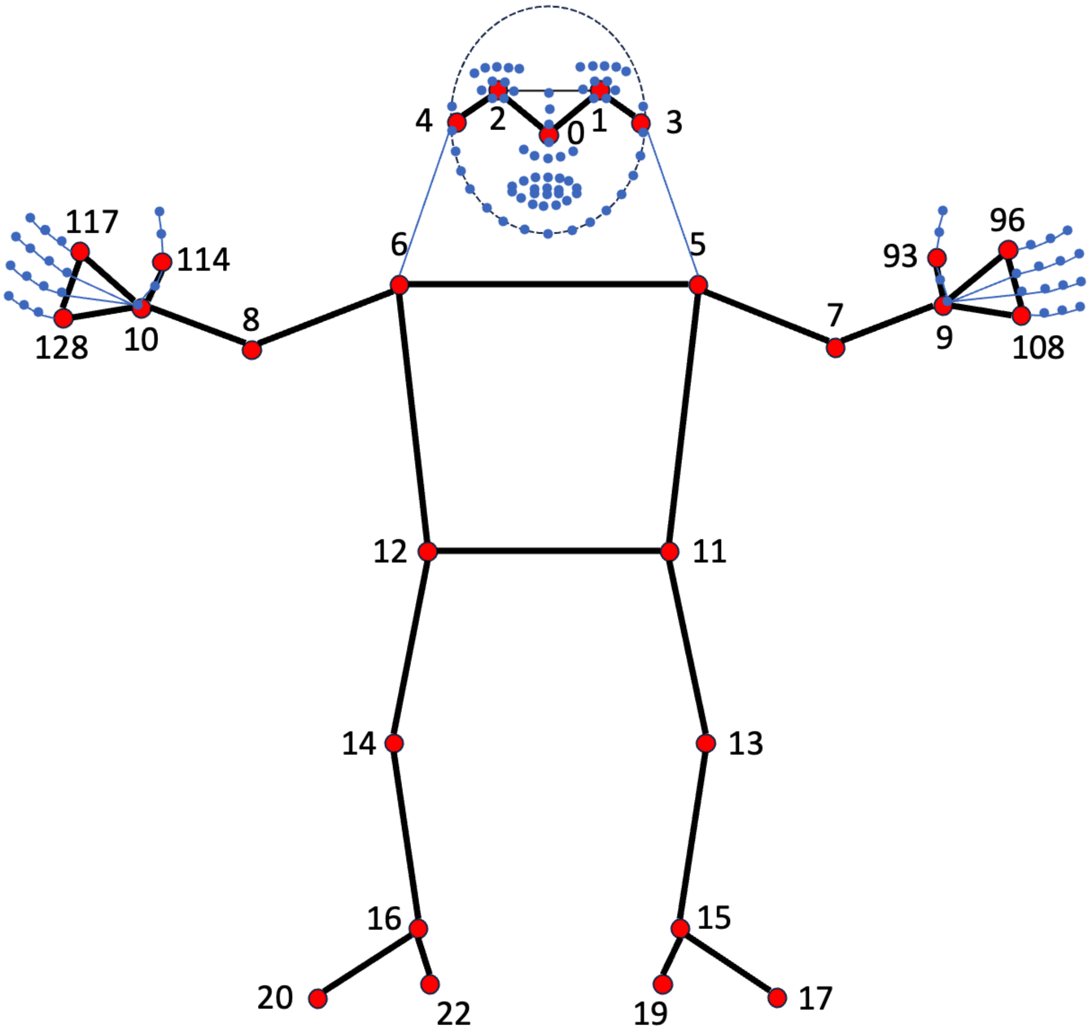

Figure 2.6
MMPose skeleton structure with IDs of used keypoint in the further processing. For simplicity, the small blue points do not have ID ensuring good visibility.

2.7 MMPose

The ouput structure of the *MMPose* model can be found in **Figure 2.6**.

2.8 Chapter Summary

This chapter introduced the key concepts of NNs, CNNs, RCNNs, transformation models, and existing models for human pose estimation. These concepts provide the theoretical foundation for the proposed automated NN learning approach.

3 Practical part

3.1 Dataset

3.2 Created Unified Format

The *unified format* structure can be found in **Figure 3.1**.

3.3 Experiments and Results

3.4 Implementation Problems and Technical Limitations

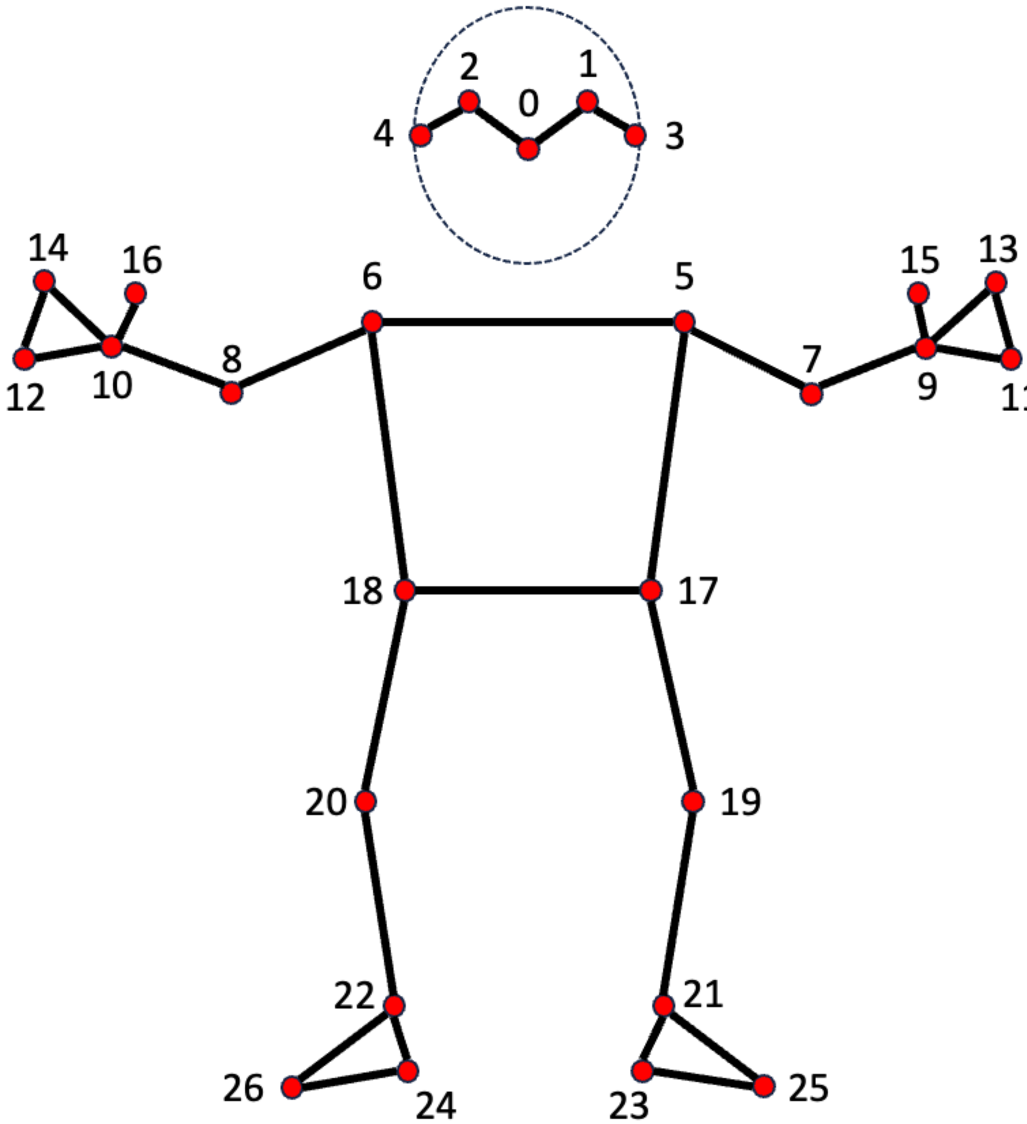

Figure 3.1
Unified format structure with IDs to each keypoint

4 Conclusion

Evaluation of the achieved results
Suggestions for further improvements
Summary of the results of the work

References

OUYANG WANLI, CHU XIAO, WANG XIAOGANG Multi-source Deep Learning for Human Pose Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* [on-line!]. 2014 [cit. 2024-01-24]. Available at: https://ieeexplore.ieee.org/document/6909696.

HE KAIMING, ZHANG XIANGYU, REN SHAOQING *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.* New York City : Springer International Publishing, 2014. [cit. 2024-01-29]. x pp. ISBN 9783319105789. Available at: http://dx.doi.org/10.1007/978-3-319-10578-9_23. DOI: 10.1007/978-3-319-10578-9_23.

GIRSHICK ROSS, DONAHUE JEFF, DARRELL TREVOR Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* [on-line!]. 2016 [cit. 2024-01-29]. Available at: https://ieeexplore.ieee.org/document/7112511.

KOUSHIK AHMED Understanding Convolutional Neural Networks (CNNs) in Depth. In *Medium* [on-line!]. Nov 2023 [cit. 2024-01-29]. Available at: https://medium.com/@koushikkushal95/understanding-convolutional-neural-networks-cnns-in-depth-d18e299bb438.

REN SHAOQING, HE KAIMING, GIRSHICK ROSS Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems* [on-line!]. 2015 [cit. 2024-01-29]. Available at: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

GOODFELLOW IAN, BENGIO YOSHUA, COURVILLE AARON *Deep Learning* [on-line!]. Cambridge : MIT Press, 2016. [cit. 2024-01-22]. 800 pp. Available at: http://www.deeplearningbook.org.

AGARWAL SHRUTI, NAGRATH PREETI, SAXENA ANMOL Human Pose Estimation Using Convolutional Neural Networks. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* [on-line!]. Feb 2019 [cit. 2024-01-28]. Available at: https://www.researchgate.net/publication/332760454_Human_Pose_Estimation_Using_Convolutional_Neural_Networks.

HUANG LIANGHUA, SHI ZHIMIN, WANG YUNHONG *Skeleton-Based Human Pose Estimation: A Survey* [on-line!]. 1. ed. New York, NY : ACM, 2022. [cit. 2024-01-20]. 39 pp. ISBN 978-1-4503-9109-3. Available at: https://dl.acm.org/doi/abs/10.1145/3541930.

Singh Anubhav, Agarwal Shruti, Nagrath Preeti Human Pose Estimation Using Convolutional Neural Networks. In *2019 Amity International Conference on Artificial Intelligence (AICAI)* [on-line!]. New York City : IEEE, 2019 [cit. 2024-1-24]. Available at: https://ieeexplore.ieee.org/document/8701267.

Mazur Matt Backpropagation in Neural Networks: An Introduction. In *mattmazu* [on-line!]. march 2015 [cit. 2024-01-20]. Available at: https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/.

Nielsen Michael A. *Neural Networks and Deep Learning* [on-line!]. Online : Determination Press, 2015. [cit. 2024-01-22]. unknown pp. Available at: http://neuralnetworksanddeeplearning.com/.

Szeliski Richard *Computer Vision: Algorithms and Applications* [on-line!]. 1. ed. London : Springer, 2010. [cit. 2024-1-23]. 812 pp. ISBN 978-1-84882-935-0. Available at: https://szeliski.org/Book/1stEdition.htm.

Simoneau Matthew J., Price Jane Neural Networks Provide Solutions to Real-World Problems: Powerful new algorithms to explore, classify, and identify patterns in data. In *MathWorks* [on-line!]. 1998 [cit. 2024-01-22]. Available at: https://www.mathworks.com/company/newsletters/articles/neural-networks-provide-solutions-to-real-world-problems-powerful-new-algorithms-to-explore-classify-and-identify-patterns-in-data.html.

Toshev Alexander, Szegedy Christian DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* [on-line!]. New York City : IEEE, June 2014 [cit. 2024-01-23]. Available at: http://dx.doi.org/10.1109/CVPR.2014.214.

Yang Wei, Ouyang Wanli, Wang Xiaolong 3D Human Pose Estimation in the Wild by Adversarial Learning. In *researchgate* [on-line!]. March 2018 [cit. 2024-01-23]. Available at: https://www.researchgate.net/publication/324055538_3D_Human_Pose_Estimation_in_the_Wild_by_Adversarial_Learning.

List of Tables

List of Figures

2.1	Example neural network schema. Source: (Nielsen, 2015)	12
2.2	A simple classification architecture by CNN. Source: (Koushik, 2023)	13
2.3	RCNN stages. Source: (Girshick, 2016)	16
2.4	Posenet skeleton structure with IDs to each keypoint	17
2.5	Movenet skeleton structure with IDs to each keypoint	18
2.6	MMPose skeleton structure with IDs of used keypoint in the further processing. For simplicity, the small blue points do not have ID ensuring good visibility.	19
3.1	Unified format structure with IDs to each keypoint	22

List of Abbreviations

CNN	Convolutional neural network
NN	Neural network
RCNN	Region-based convolutional neural network

List of Source Codes

APPENDICES