BABEŞ BOLYAI UNIVERSITY CLUJ NAPOCA

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

SPECIALIZATION SOFTWARE ENGINEERING

DISSERTATION THESIS

Real-time human pose estimation and tracking to record the progress of rehabilitation therapy

Supervisor

Lect. Dr. Ioan Lazar

Author

Razvan Timis

2019

Table of Contents

Abstract

The goal of this paper is to obtain reports on patient progress based on intelligent pose estimation algorithms motivate the patient to continue medical recovery.

Most patients require long-term medical recovery, which means an average of 2 or 3 years, give up during the recovery period because they do not see results even if they exist, they are not obvious to the naked eye.

More worrying is that up to 70% of patients give up physiotherapy because they can not see immediate results. [7].

Our solution is an application called My Kineto through which the patient will be able to record their therapy sessions and will receive constant feedback in real time about Range of Motion. Using this data these data, we will generate reports on patient progress.

Chapter 1

Introduction

In Physiotherapy, tracking Range of Motion (ROM) is a standard approach to measuring progress in patient therapy. Often, ROM is measured subjectively and documentation is inconsistent between clinicians. Physios might come to wrong conclusions if ROM is tracked incorrectly between therapy sessions.

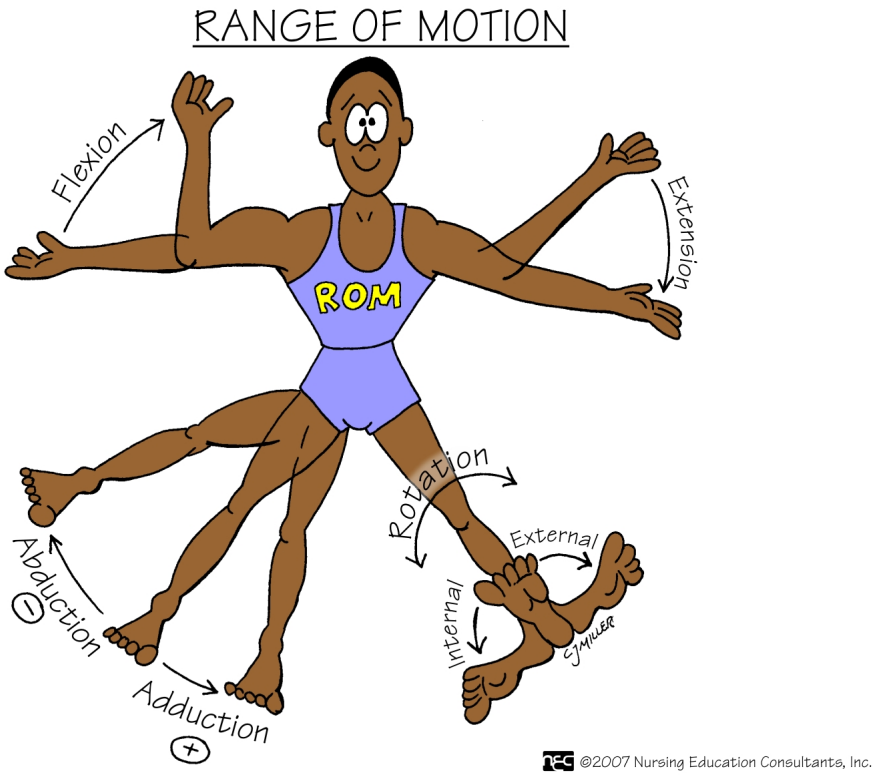

Figure 1.1: Range of Motion

The problem is that up to 70% of patients give up physiotherapy because they can not see imme-

diate results [7].

That's why we want to make a application which makes use of a camera to objectively calculate ROM in real-time and automatically produce a report that tracks progress over the course of several therapy sessions.

Chapter 2

Theoretical Background

In this chapter we will present a brief introduction to neural networks and then describe the underlying concepts of the Convolutional Neural Network (CNN). We will focus on presenting the main concepts that will be used in the development of pose estimation algorithms.

2.1 Neural network

In the Neural Networks chapter, we will talk in the first part about a brief history, presenting the three periods through which artificial intelligence has passed. Then we will address the biological part that is found in neural networks. In the next section we will discuss what is a neural network , how many types of such networks are there? What is a neuron? But an activation function? Finally, we will discuss learning such a neural network with its specific algorithms.

2.1.1 History

Artificial intelligence takes us to think of some SF films, but it still has a long way to go, for now these intelligent algorithms can only get to the level of intelligence of an insect, they work better in certain exact tasks as imagine detectie and not in general like a brain.

But let's not forget that this domain has been built around the dream to overcome human intelligence. The essential question is whether such a system can be implemented on a computer?

The domain of psychology was the first to have had the artificial intelligence applicability, the most famous Turing test that appeared in 1950. It involves a conversation of a person with a computer

and another person and he has to guess who is the person. [14]

Three great periods are in the history of artificial intelligence. In the first period, only after the Second World War. The first programs that implement various smart algorithms to solve puzzles. [14]

An important algorithm in this field is Samuels' game, it was quite simple to implement. They save certain winning positions throughout the game.[14] The first period kept until 1965, but no algorithm has led to major changes in people's lives. [14]

In the second period, it focuses on the processing of natural language so many applications are launched that implement concepts from the processing of natural language.

One of the famous programs of those times was called ELIZA. This program learns to copy the conversations of a psychologist with his patients. How did ELIZA work? ELIZA had a knowledge base in English and a field of psychology made up of a set of rules, so ELIZA worked on the "fit" principle. For example, when it found the word "father," it said, "Tell me about your family" [14]

Also during this period, rule-based expert systems appeared. One such system was called MYCIN, which was designed to diagnose infectious diseases of the blood and recommend medical treatment. It is based on the rules made by specialists. [14]

From 1975 to the present, there is the third period of artificial intelligence. This domain is becoming more stable and the industry adopting these intelligent systems.

Warren McCulloch and Walter Pitts have been the first research that has made the research. In 1940, they highlighted the first digital model of a neuron that discovers the computational capacity, also providing a mathematical abstraction of this concept. Thus the synapses that a neuron makes through dendrites become inputs, the body of a neuron becomes an activation function, the axon has become the output and the BAIS notion has been introduced for mathematical restlessness features. [2]

In 1969, authors Marvin Minsky and Seymor Papert published the book "Perceptrons," which highlighted the limitations that exist for one-level neural networks. After this publication many people who doing research was quit. [2]

Neural networks have become one of the most used in our days. Big Cloud companies offer an API through which any developer can get his own training in the easiest way.

2.1.2 Biological

The great mystery of this universe is the way people think, knowing for several thousand years that powerful head shots can generate loss of consciousness or even death. But more than that, we know our brain is different from animals. In about 335 BC, Aristotle wrote, "Of all animals, man has the greatest brain." [11]

The nervous system is made up of neurons. Each neuron can be represented as a unit. Between neurons there are synapses that can be of two axo-somatic or axo-dendritic types. The body of the neuron has two types of extensions [5]:

- Dentrites are relatively short, and branched near the cell
- The axon is longer and thicker in the propagation of the electrical impulse.

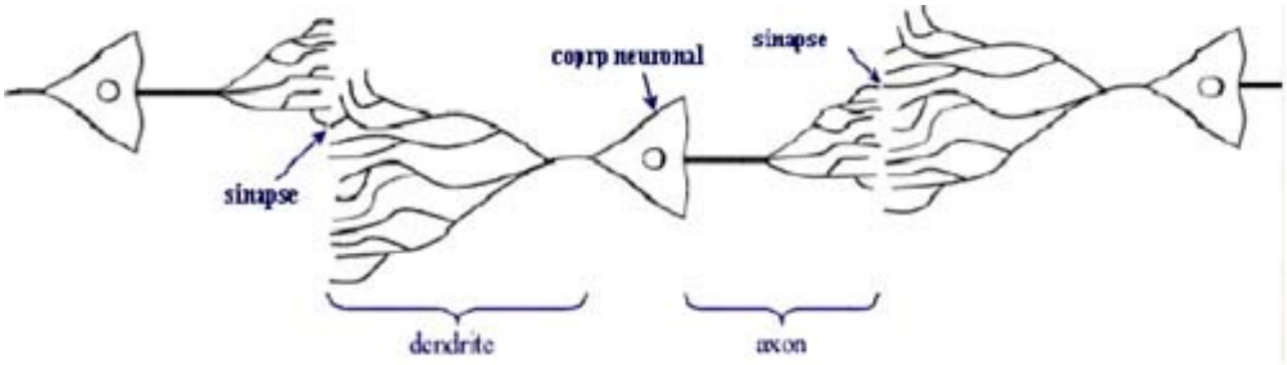

Figure 2.1: Network of biological neurons

While a reader uses a network of 10^{11} biological neurons and 10^{4} connections between neurons, the processing time of 10^{-3} ms and 10^{-9} ms for electronic circuits, although biological neurons are more slower than electronic circuits, biological ones can process information much faster than any other circuit. [5]

2.1.3 Overview

A neural network is a copy of the biological model, so it has nodes that are connected by links, these links are actually real numbers that represent the memory of a network. Learning a neural network is accomplished by changing the weight between the nodes.

Any neural network is organized on layers, so there is an input layer and an output layer. The input layer is the one that receives the data from the outside and the output layer provides us with the information we expect. [11]

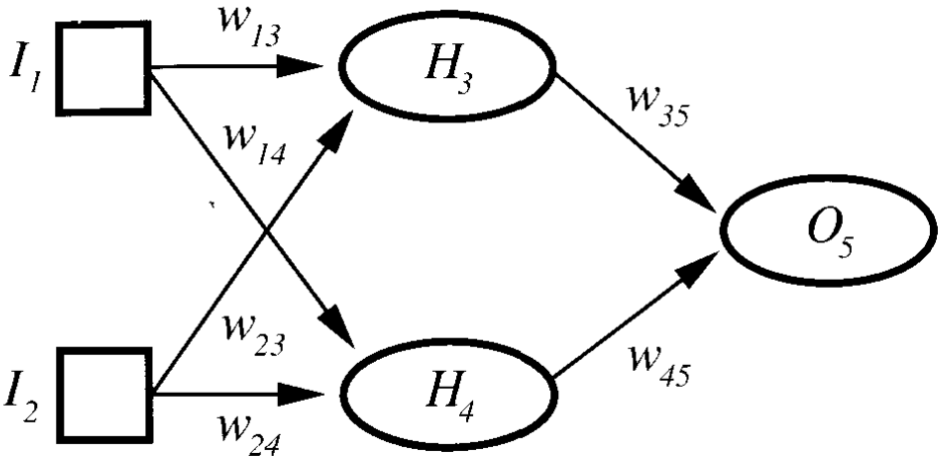

Figure 2.2: Feedforward neural network [11]

In Figure 2.2 shows a neural network of the three-layer feed-forward type, which has links between the neurons in one direction, because the cycles in such a network are missing, the calculation can continue evenly from the input nodes to the exit ones [11].

2.1.4 Back propagation algorithm

2.2 Convolutional Neural Network

2.2.1 What is differente?

Chapter 3

Related work

3.1 DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model [6]

3.1.1 Problem

The goal of this paper is to advance the state-of-the-art of articulated pose estimation in scenes with multiple people.

3.1.2 Methods

Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. The architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving realtime performance.

The part affinity is a 2D vector field for each limb: for each pixel in the area belonging to a particular limb, a 2D vector encodes the direction that points from one part of the limb to the other. Each type of limb has a corresponding affinity field joining its two associated body parts.

3.1.3 Data

The MPII human multi-person dataset [3] and the COCO 2016 keypoints challenge dataset [9]. These two datasets collect images in diverse scenarios that contain many real-world challenges such as crowding, scale variation, occlusion, and contact.

MPII Human Pose dataset is a state of the art benchmark for evaluation of articulated human pose estimation. The dataset includes around 25K images containing over 40K people with annotated body joints. The images were systematically collected using an established taxonomy of everyday human activities.

3.1.4 Performance and comparisons

Evaluation is done on two single-person and two multi-person pose estimation benchmarks. The proposed approach significantly outperforms best known multi-person pose estimation results while demonstrating competitive performance on the task of single person pose estimation.

3.2 Human Pose Estimation from Monocular Images: A Comprehensive Survey [15]

3.2.1 Problem

In this paper, a comprehensive survey of human pose estimation from monocular images is carried out including milestone works and recent advancements. The goal of our application is to provide initialization for automatic video surveillance.

3.2.2 Methods

Based on one standard pipeline for the solution of computer vision problems, this survey splits the problem into several modules: feature extraction and description, human body models, and modeling methods. There are additional sections for motion-related methods in all modules: motion features, motion models, and motion-based methods.

3.2.3 Data

The paper collects 26 publicly available datasets for validation and provides error measurement methods that are frequently used.

3.2.4 Performance and comparisons

The first survey that includes recent advancements on human pose estimation based on deep learning algorithms. Although deep learning algorithms bring huge success to many computer vision problems, there are no human pose estimation reviews that discuss these works. In this survey, about 20 papers of this category are included. This is not a very large number compared to other problems, but this is a inclusive survey considering the relatively few works addressing this problem.

3.3 PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model [12]

3.3.1 Problem

Paper try to solve the task of 2-D pose estimation and instance segmentation of people in multi-person images using an efficient single-shot model based on a box-free bottom-up approach. The study contributes to computer vision applications such as smart photo editing, person and activity recognition, virtual or augmented reality, and robotics.

3.3.2 Methods

Bottom-up approach for pose estimation and segmentation:

- localizing identity-free semantic entities (individual keypoint proposals or semantic person segmentation labels, respectively)
- grouping them into person instances:
  - use a greedy decoding process to group them into instances
  - train network to predict instance-agnostic semantic person segmentation maps
  - for every person pixel we also predict a set vectors to each of the K keypoints of the corresponding person instance. (corresponding vectors fields can be thought as a geometric

embedding representation and induce basins of attraction around each person instance, leading to an efficient association algorithm)

  * For each pixel xi, they predict the locations of all K keypoints for the corresponding person that xi belongs to
  * then compare this to all candidate detected people j (in terms of average keypoint distance), weighted by the keypoint detection probability
  * if this distance is low enough, we assign pixel i to person j

3.3.3 Data

Standard COCO keypoint dataset [9] , which annotates multiple people with 12 body and 5 facial keypoints.

3.3.4 Performance and comparisons

- compared to the best previous bottom-up approach they improve keypoint AP (Average Precision) from 0.655 to 0.687
- compared to the strong top-down FCIS method [8] improve mask AP from 0.417 to 0.386

3.4 Towards Accurate Multi-person Pose Estimation in the Wild [13]

3.4.1 Problem

Paper try to solve the task of multi-person detection and 2D pose estimation based on a top-down approach (2-D localization of human joints on the arms, legs, and keypoints on torso and the face). The study want to localize people, understand the activities they are involved in, understand how people move for the purpose of Virtual/Augmented Reality, and learn from them to teach autonomous systems.

3.4.2 Methods

Top-down approach consisting of two stages:

- first stage, predict the location and scale of boxes which are likely to contain people, use Faster RCNN detector.
- second stage, estimate the keypoints of the person potentially contained in each proposed bounding box:
  - for each keypoint type predict dense heatmaps and offsets using a fully convolutional ResNet
  - combine these outputs by a novel aggregation procedure to obtain highly localized keypoint predictions
    * use keypoint-based Non-Maximum-Suppression (NMS), instead of the cruder boxlevel NMS
    * keypoint-based confidence score estimation, instead of box-level scoring

3.4.3 Data

Training data: standard COCO keypoint dataset [9], which annotates multiple people with 12 body and 5 facial keypoints.

3.4.4 Performance and comparisons

- average precision of 0.649 on the COCO test-dev set and the 0.643 test-standard sets, outperforming the winner of the 2016 COCO keypoints challenge and other recent state-of-art
- using additional in-house labeled data we obtain an even higher average precision of 0.685 on the test-dev set and 0.673 on the test-standard set, more than 5% absolute improvement compared to the previous best performing method on the same dataset

3.5 Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [4]

3.5.1 Problem

The purpose is to recognize a layout of the body parts, based on the pose estimation and Pose Affinity Fields (PAF)

3.5.2 Methods

- Part affinity fields(PAF): is a set of 2D vector fields that encode the location and orientation of the limbs. It is a set of vectors that encodes the direction from one part of the limb to the other; each limb is considered as an affinity field between body parts.
- Estimation of body-part confidence maps.

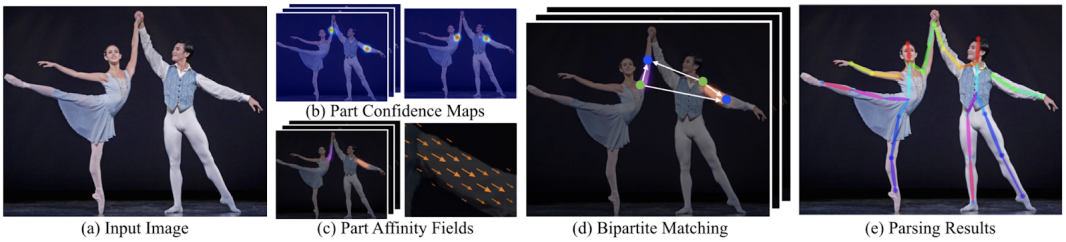

Figure 3.1: Overall pipeline

3.5.3 Data

As an input we have an image. This method simultaneously infers two maps with body-parts, as you can see in b figure and runs a special bipartite matching algorithm. In the end it assemble the body parts into full body poses.

Training data: standard COCO keypoint dataset [9] and the MPII human multi-person dataset [3]

Chapter 4

MyKineto - Application

My Kineto application is an interactive software destined for patients who need physiotherapy treatments. Our application guide patient to see how they need to make their exercises in a correct manner by showing Range of Motion (ROM) in real time and after allowing us to report on these dates for motivation patient. Also, the application count the number of movements. It is based on exercices, because we think that a constant and correct number of movements could be more efficient for patients then just to present them what exercices they need to do. In this manner the application guides each patient during the all period they need to follow their treatment.

4.1 Motivation

Our motivation is to help patients who need physiotherapy to see a progress and to encourage them to be constant during their treatment. We want to support patients motivation in continuing to build new and healthy behaviours.

This application is designed to help everyone who need physiotherapy treatment to stay motivated, reach their goals, and create habits that are healthy and helpful for a long term. In this case we have a solution by creating a app which will be a useful tool for patients who needs help to reach their goals by automated tracking of ROM.

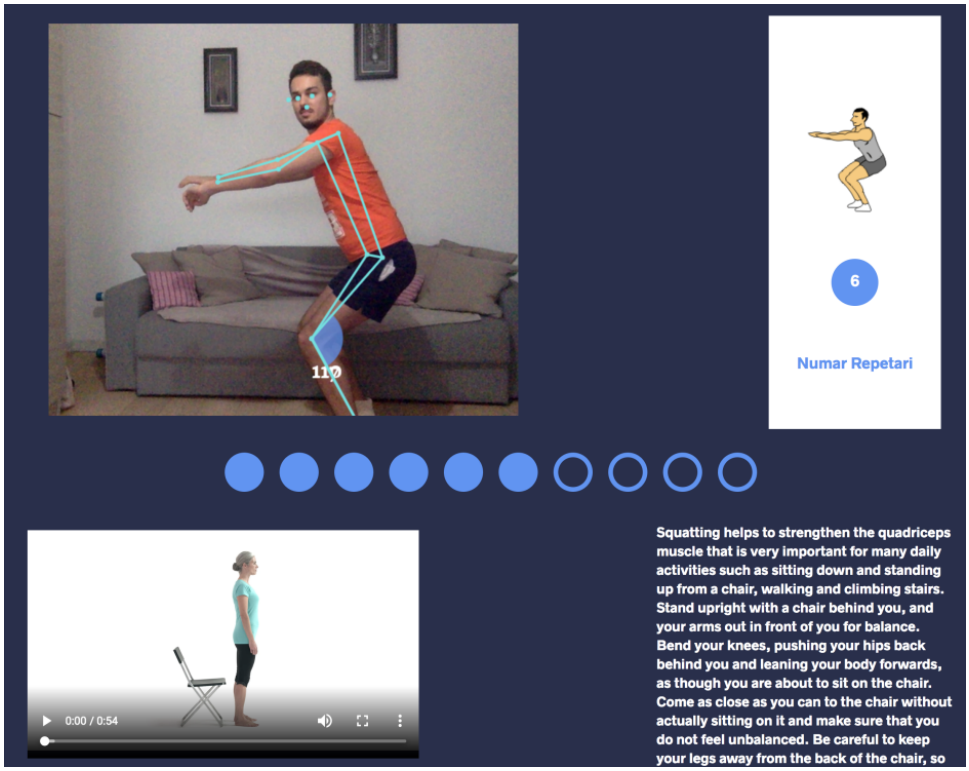

Figure 4.1: Demo of My Kineto

4.2 Application development

4.2.1 Specification of the problem

4.2.2 Analysis and design

4.2.3 Implementation

4.2.4 Posture tracking algorithm

4.2.4.1 Initialization

Initialization is the first part of the pose tracking algorithm, where we detect the posture with posenet and for each part of the detected skelet we calculate a bounding box and detect the initial set of keypoints that we are going to track in the next steps.

4.2.4.2 Processing

Processing is composed from three main parts:

- Tracking
- Add new points
- Respawn / Reinitialization

This part is used to track as much as possible points detected in the first phase.

4.2.4.3 Keypoints tracking

Points detected in the first part will be used inside tracking algorithm Lucas-Kanade [10] (function calcOpticalFlowPyrLK() from OpenCV). This algorithm fits into the Detect Track class (DT) and makes a local search to determine the new position of the points of interest detected in the previous frame. For this algorithm to work fine, it's important to take consecutive frames that are easily modified. If we do a sudden move, this algorithm will fail to detect to detect the new position of a part or even all the points in the previous frame. For this case, the next two steps in the algorithm will try to restore the system.

The points in the system that are tracked are detected with Shi-Tomasi "Good features to track" [1].

After we detect new position for our keypoints we have to detect the new bounding box that correspond to their new position. For this task we choose to detect the homography matrix based on old points and new points position. Based on homography matrix we calculate the perspective transformation of previous bounding box.

4.2.4.4 Add new points

Adding new points is a step for restoring the system. As we track certain points, we may lose track of some of them. In this case, to prevent destabilization of the system, we will add new key points.

The system starts with a maximum number of key points. If a percentage of N of the maximum number of points is lost, then we will try to find others to replace the missing ones.

The detection of new points will be achieved with the Shi-Tomasi algorithm "Good features to track". To narrow the search space of these algorithms, their implementations in OpenCV allow the definition of a mask.

A mask is an image size matrix in which we want to find the key points. To mark the fact that we want to search in a certain area we will set the value of 1, the mask, in that region and 0 in the rest. Applying a mask is important not only to narrow the search space but also to don't keep the points detected in an area where our posture skeleton is not found.

More specifically in our application, for each bounding box built from the skelet we will add new points inside it. The calculated bounding box is used to create a mask that we can use inside the keypoints detection algorithm.

4.2.4.5 Respawn / Reinitialization

System reinitialization consists in the identification of the fact that we lost almost all of the tracked points and we are no longer able to estimate, with the remaining number of points, the skeleton posture. The algorithms used to reinitialize the posture are the ones used in the first phase. So basically from this phase if we are no longer able to track and estimate new posture of our skeleton then we go the the first step (Initialization).

System reinitialization denotes that the system has been completely destabilized. Loss of a large number of points can occur either due to a sudden movement or because the tracking posture is no longer in the frame.

4.2.5 User manual

4.3 Experimental results and comparisons with similar approaches

4.4 Possible extensions

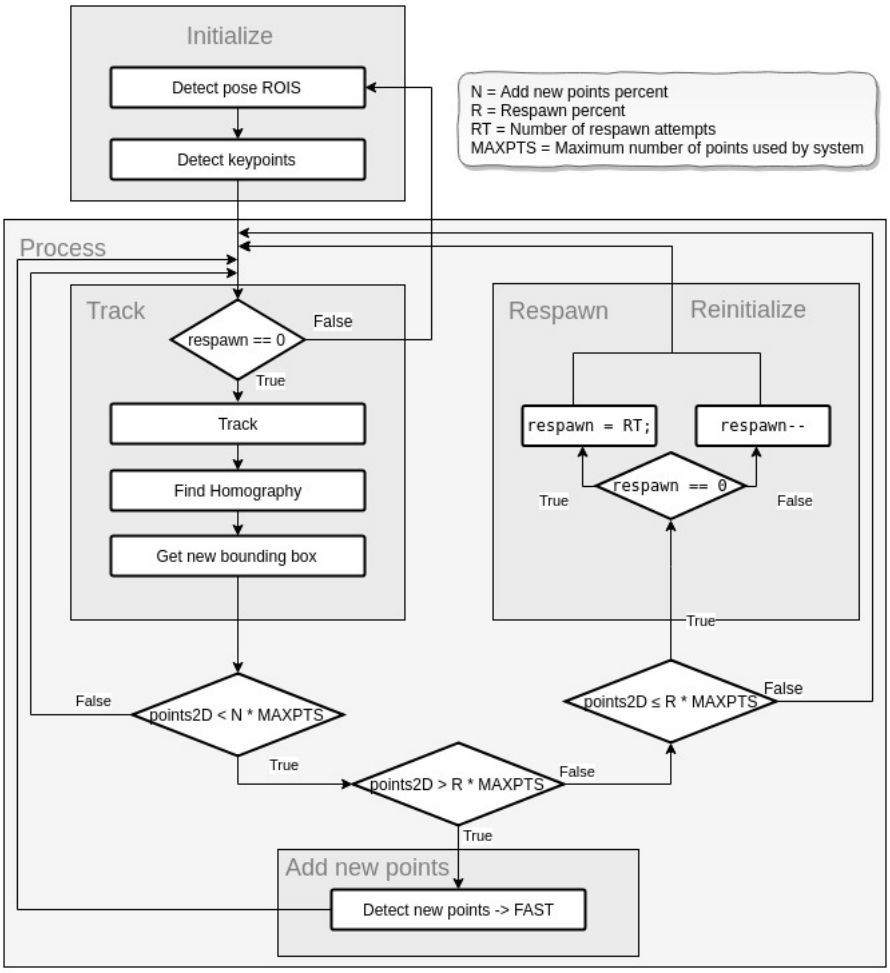

Figure 4.2: Posture tracking algorithm

Chapter 5

Conclusion

Proposed system for the tracking of the posture doesn't perform as we originally expected.

There seems to be some problems regarding the determination of the bounding box, maybe due to the fact that the detected points are on a surface that is characterized by just a few colors that appear in almost all tracked region. Because of these characteristics of the tracked posture we fail to track enough points due to insufficient difference in the nearest proximity of a point.

For future development we will try to adapt described system to track the contours.

Even if posture tracking system doesn't perform as we expected, we succeeded in the finding the appropriate settings for the PoseNet so that it performs enough good for the determination of the ranges of motion. Based on these calculations we were able to count, in real time, the number of exercises that user made.

Bibliography

[1] Proceedings of ieee conference on computer vision and pattern recognition. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 1994. doi: 10.1109/CVPR.1994.323798.

[2] Neuronal networks - history. https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/, May 2019.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. 2014. In CVPR.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016.

[5] Conf. dr. Mircea Ifrim, Prof. dr. doc. Gherghe Niculescu, Prof. dr. N. Bareliuc, and Dr. B. Cerbulescu. *Atlas de anatomie umana, Volumul III*. 1985.

[6] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *CoRR*, abs/1605.03170, 2016.

[7] Ryan Klepps. 7 thought-provoking facts about physical therapy you can't ignore. https://www.webpt.com/blog/post/7-thought-provoking-facts-about-physical-therapy-you-cant-ignore, May 2019.

[8] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. *CoRR*, abs/1611.07709, 2016.

[9] Lin, T.Y., Cui, Y., Patterson, Bourdev, Girshick, and Dollr. Coco 2016 keypoint challenge. 2016.

[10] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[11] Peter Norvig and Stuart J. Russell. *Artificial Intelligence - A Modern Approach*. Prentice Hall, 1994.

[12] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *CoRR*, abs/1803.08225, 2018.

[13] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin P. Murphy. Towards accurate multi-person pose estimation in the wild. *CoRR*, abs/1701.01779, 2017.

[14] Raluca Vasilescu. Un scurt istoric al ia. https://www.cs.cmu.edu/~mihaib/articole/ai/ai-html.html, May 2019.

[15] Gong W, Zhang X, GonzÃ lez J, Sobral A, Bouwmans T, Tu C, and Zahzah EH. Human pose estimation from monocular images: A comprehensive survey. *PubMed*, PMID: 27898003, 2016. doi: 10.3390/s16121966.