FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Computer Vision for Accessible Rehabilitation

José Miguel Afonso Maçães

Mestrado em Engenharia Informática e Computação

Supervisor: Ivone Amorim

Second Supervisor: Bruno Cunha

February 16, 2024

Computer Vision for Accessible Rehabilitation

José Miguel Afonso Maçães

Mestrado em Engenharia Informática e Computação

February 16, 2024

Abstract

Physical rehabilitation is vital to injury recovery, providing pain relief and shortening recovery time. However, traditional rehabilitation methods heavily rely on in-person professional feedback, which can be time-consuming, expensive, and prone to human error. This often limits the accessibility and effectiveness of rehabilitation services, which is why patients are suggested to perform rehabilitation exercises at home. Yet, patients become demotivated without professional feedback and do not follow through with rehabilitation.

Accessible rehabilitation is in demand and there is a possibility of solving this using something everyone has access to: a smartphone. This would allow patients to receive feedback on exercises independently via a system that, fed as input a video of the patient performing the exercise, outputs specific feedback as a measure of how correctly the exercise is being performed. Succinctly, the goal would be to assess human motion quality in rehabilitation exercises without in-person supervision or special equipment.

This work reviews the current state of the art in Computer Vision systems for assessing physical rehabilitation exercises and develops and tests an intelligent system that can aid patients in their rehabilitation process.

More specifically, this work proposes a motion tracking system based on Computer Vision that receives as input a video of a rehabilitation exercise recorded with a smartphone. The system prioritizes accessibility, opting for the use of accessible technology, and is evaluated by comparing it to a state-of-the-art complex rehabilitation assessment system, the Qualysis Motion Capture System, based on a labeled dataset curated by expert physicians. The key points of the proposed framework are the human pose detection method and the human movement quality assessment.

The proposed system has the potential to reduce recovery times, eliminate human error, and make rehabilitation more accessible.

Resumo

A reabilitação física é vital para a recuperação de lesões, proporcionando alívio da dor e reduzindo o tempo de recuperação. No entanto, os métodos tradicionais de reabilitação dependem fortemente do feedback profissional presencial, que pode ser demorado, caro e sujeito a erros humanos. Isto limita a acessibilidade e eficácia dos serviços de reabilitação, razão pela qual os pacientes são aconselhados a realizar exercícios de reabilitação em casa. No entanto, os pacientes ficam desmotivados sem o feedback profissional e não seguem adiante com a reabilitação.

A reabilitação acessível está em falta e há a possibilidade de resolver isso usando algo a que todos têm acesso: um smartphone. Isto permitiria que os pacientes recebessem feedback sobre os exercícios de forma independente através de um sistema que, ao receber como entrada um vídeo do paciente executando o exercício, fornecesse feedback específico indicando quão corretamente o exercício é realizado. Sucintamente, o objetivo seria avaliar a qualidade do movimento humano em exercícios de reabilitação sem supervisão presencial ou equipamento especial.

Este trabalho inclui uma revisão do estado da arte de sistemas de Visão por Computador para avaliar exercícios de reabilitação física e desenvolve e testa um sistema inteligente que pode ajudar os pacientes no seu processo de reabilitação.

Mais especificamente, este trabalho propõe um sistema de *motion tracking* baseado em Visão por Computador que recebe como input um vídeo de um exercício de reabilitação gravado com um smartphone. O sistema dá prioridade à acessibilidade, optando pelo uso de tecnologia acessível, e é avaliado comparando-o a um sistema de avaliação de reabilitação de alta qualidade, o *Qualysis Motion Capture System*, com base num conjunto de dados avaliado por médicos especialistas. Os pontos-chave da framework proposta são o método de detecção de pose humana e a avaliação da qualidade do movimento humano.

O sistema proposto tem o potencial de reduzir os tempos de recuperação, eliminar erros humanos e tornar a reabilitação mais acessível.

Agradecimentos

First of all, I thank my supervisors, Professor Ivone Amorim and Professor Bruno Cunha for their outstanding support throughout this entire journey. Their constant availability and motivation were key to the success of this project.

I thank my family, who are always there for me, in the good and in the bad. Their unwavering support and belief in me have undoubtedly shaped me as a student, but most importantly as a human.

I also thank my friends, who put up with me during this entire process and held me up whenever I needed them to.

José Miguel Afonso Maçães

"Viajar?
Para viajar basta existir."

Fernando Pessoa

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
CC	Correlation Coefficient
CNN	Convolutional Neural Network
CPM	Convolutional Pose Machines
CV	Computer Vision
DL	Deep Learning
DTW	Dynamic Time-Warping
ESS	Escola Superior de Saúde
GAN	Generative Adversarial Network
GE	Guided Exercises
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HSMM	Hidden Semi-Markov Model
IPP	Instituto Politécnico do Porto
MAE	Mean Absolute Error
NGE	Non-Guided Exercises
NN	Neural Network
PORTIC	Porto Research, Technology and Innovation Center
QTM	Qualisys Motion Capture System
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
ROM	Range of Motion
SDK	Software Development Kit
VR	Virtual Reality

Chapter 1

Introduction

Rehabilitation plays a vital role in injury recovery and in restoring the physical capabilities of individuals who have disabilities or other health conditions. Traditional rehabilitation approaches often rely on in-person sessions supervised by healthcare professionals, which can be costly, time-consuming, and may lack accessibility for certain populations [39]. In recent years, there has been a growing interest in integrating advanced technologies to augment rehabilitation processes and improve patient outcomes.

One area of technology that holds significant promise in the field of rehabilitation is Computer Vision (CV). CV allows computers to interpret visual data through image processing and machine learning algorithms. By capturing and analyzing the movements of individuals, these systems should be able to provide feedback without the need for the availability of a healthcare professional. They have the potential to revolutionize rehabilitation by enabling increased access to rehabilitation services.

In this thesis, we focus on the application of CV for accessible physical rehabilitation. To accomplish this, we will explore the use of markerless motion tracking techniques, which eliminate the need for external sensors or physical markers. We will investigate various components of these CV systems and propose a solution.

Overall, this thesis aims to contribute to the growing body of knowledge in the field of CV systems for rehabilitation by developing a motion tracking system that can improve the accessibility of rehabilitation processes. Through our work, we hope to empower individuals to take control of their rehabilitation journey, enabling them to enhance their overall well-being.

1.1 Context

To better understand the work that will be done, this section will focus on giving context and some general concepts, namely regarding rehabilitation.

1.1.1 Rehabilitation

According to the World Health Organization [60], rehabilitation refers to a set of interventions designed to enhance functioning and reduce disability in individuals with health conditions in interaction with their environment. It is the process of restoring a person's physical, cognitive, or psychological abilities after an illness, injury, or surgery. It involves a person-centered approach to help individuals regain functional independence, mobility, and quality of life. The aim is to address impairments, disabilities, and limitations through various therapeutic interventions tailored to each patient's needs. Ultimately, via rehabilitation, people can overcome difficulties with thought, sight, hearing, communication, or moving around.

Despite the specific goals and methods of rehabilitation varying depending on the individual's condition, common objectives include reducing pain, improving strength and flexibility, enhancing mobility, and restoring cognitive function.

Rehabilitation plays a vital role in reducing the impact of various health conditions, including acute and chronic diseases, illnesses, and injuries. It serves as a complement to medical and surgical interventions, facilitating recovery and optimizing overall outcomes. Rehabilitation helps minimize or slow down the disabling effects of chronic health conditions, promoting healthy aging. It also aids in avoiding costly hospitalizations, reducing hospital stays, preventing readmissions, and reducing the need for financial support and caregiver assistance.

Yet, the global demand for rehabilitation remains largely unfulfilled due to several factors. These include the limited availability of rehabilitation services in some areas, leading to long waiting times for individuals in need. Furthermore, there is a lack of trained rehabilitation professionals, with many settings having less than ten skilled practitioners per 1 million people. Another challenge is insufficient access to assistive technology and equipment [60].

As such, we can state that the main problem with conventional rehabilitation methods is accessibility, as justified in the previous paragraph. One attempt to solve this involves prescribing rehabilitation exercises to the patient and instructing them to perform the mentioned exercises at home [29], where they are not necessarily bound by a healthcare professional's schedule or availability. This process can also be supported by telerehabilitation, which is explained in the next section. In spite of that, the most common outcome of this home-based setup is patient demotivation and non-compliance. This is primarily due to the absence of direct supervision and real-time feedback from professionals [4].

1.1.2 Telerehabilitation

Telerehabilitation is the use of information and communication technology to deliver or support the delivery of rehabilitation [48]. This can encompass various rehabilitation services, such as evaluation, assessment, monitoring, prevention, intervention, supervision, education, consultation, and coaching. The use of information and communication technologies to deliver these services can include video and audio conferencing, messaging via chat, wearable technologies, sensor technologies, patient portals or platforms, mobile health applications, Virtual Reality (VR), robotics, and therapeutic gaming technologies [48]. These services are delivered by professionals, sometimes with the help of paraprofessionals, family members, or caregivers of the patient [48].

Although telerehabilitation has proven to be a valid approach, it still faces challenges related to the availability of healthcare professionals, as the dependency on physicians for guidance and feedback on rehabilitation exercises remains a significant limitation in terms of accessibility. However, the integration of intelligent technologies like machine learning and Artificial Intelligence (AI) presents opportunities to address this issue.

By integrating CV, Deep Learning (DL), and AI algorithms, intelligent systems can provide real-time feedback to patients performing rehabilitation exercises remotely. These technologies have the potential to assess the quality of motion, provide the patient with such assessment, and offer personalized recommendations or corrections. Such systems can do all this without constant professional supervision, reducing the need for physician availability and making rehabilitation more accessible, efficient, and cost-effective. Furthermore, adherence to at-home rehabilitation could grow.

1.1.3 Intelligent Rehabilitation

Intelligent rehabilitation is presented here as a variant of telerehabilitation, but one that focuses on using more advanced or AI-based technology to aid in the delivery of rehabilitation. There have been various attempts to make use of different intelligent technologies for rehabilitation, such as robotics, VR, and CV, to enhance the rehabilitation process. Next is an overview of these approaches and their potential applications in intelligent rehabilitation. These approaches have been reviewed in [27].

Robot-assisted rehabilitation systems involve the use of robotic devices to assist patients in performing therapeutic exercises. In robotic systems, the typical setup involves interaction between the patient and the robot, in which the robot facilitates the repetition of certain movements with precision. A review of the application of robotics to rehabilitation is presented by [26]. The main drawback of this approach is its high cost, primarily due to the limited and specialized utility that these robots offer.

Following what is reported in [27], VR technology immerses users in a computer-generated environment, simulating real-world scenarios or creating virtual rehabilitation environments. VR-based rehabilitation systems provide interactive experiences that can motivate and challenge patients during their therapy sessions. VR systems enable users to interact with virtual objects and perform rehabilitative exercises in a controlled environment. This technology allows therapists to customize rehabilitation programs, monitor progress, and provide real-time feedback to patients. The primary challenge associated with VR is related to accessibility. This is because the patient is required to have special equipment to participate in the experience. Moreover, it requires the development of a dedicated virtual world specifically tailored for rehabilitation purposes.

CV emerges as the most promising option, holding significant potential for addressing the challenges at hand. In Chapters 2 and 3, we take a deeper look into this field, comprehensively explaining its principles and analyzing the innovative solutions it has presented.

1.2 Motivation

The motivation behind this research comes from recognizing the significant challenges and limitations faced by individuals who require rehabilitation services. Section 1.1 uncovers the problems with the existing approaches for rehabilitation, not only traditional approaches but also more recent ones that resort to advanced technology. Simply put, rehabilitation is not accessible enough in terms of time, money, and availability.

Developing a CV rehabilitation system aims to overcome these barriers and make rehabilitation more accessible to a wider population. This research can provide individuals with the autonomy to perform rehabilitation exercises in the comfort of their own homes, eliminating the need for extra equipment or constant supervision. This accessibility can extend to individuals with limited access to healthcare facilities or those facing financial constraints.

Overall, the motivation behind this research lies in the desire to revolutionize the field of rehabilitation by making use of accessible technology to create an effective platform that empowers individuals.

1.3 Objectives

Our objective is to develop a Computer-Vision-based system that allows patients to perform rehabilitation exercises independently, using accessible technology, while receiving necessary and adequate feedback.

1.3.1 Specific Objectives

The specific objectives of this research are as follows:

1. Review the current state of the art in Computer Vision systems for assessing physical rehabilitation exercises.
2. Develop a system capable of accurately tracking and analyzing human movements during rehabilitation exercises by comparing an example video to a reference video. This involves implementing CV algorithms to extract relevant joint positions and motion features.
3. Develop an assessment module within the system to evaluate the performance of the example video in relation to the reference video. The assessment module will analyze various aspects of the exercise execution, such as correctness and alignment with the reference.
4. Design and implement a scoring mechanism that assigns a quantitative score to the example video based on the assessment performed in Objective 3. The scoring algorithm will take into account factors such as accuracy, precision, and overall performance to provide an objective evaluation metric.

By achieving these specific objectives, the research aims to develop a system for evaluating the performance of rehabilitation exercises.

1.4 SmartHealth Project Collaboration

This master's thesis is part of a larger collaborative project called SmartHealth [46], conducted in partnership with the Porto Research, Technology and Innovation Center (PORTIC) and the Escola Superior de Saúde (ESS) at the Instituto Politécnico do Porto (IPP). The SmartHealth project aims to explore the use of technology to create more efficient and intelligent systems to support medical treatment in all of its stages.

As a result of the SmartHealth project, a prototype for an application that serves as a platform for remote rehabilitation has already been developed. This application provides a body motion capture module that allows users to perform rehabilitation exercises autonomously but currently lacks a correct assessment of the exercises. The integration of the system developed as a result of this thesis into the SmartHealth project should contribute to its further advancement.

1.5 Dissertation Structure

Besides the introduction, this thesis contains six more chapters. Chapter 2 gives background on the field of CV. In Chapter 3, the state of the art and the literature review are detailed. Chapter 4 describes the problem at hand. In Chapter 5, the implementation of the solution is described. Chapter 6 presents and analyzes results. Chapter 7 contains the conclusion and identifies the direction of future work.

Chapter 2

Background

This chapter gives some background on the topic of Computer Vision. The first section establishes the concept of CV and gives an overview of how this topic can relate to rehabilitation. Then, there is a section on Neural Networks (NNs) and Deep Learning.

2.1 Computer Vision

2.1.1 General Overview

Computer Vision is a field of AI and image processing that focuses on enabling computers to extract meaningful information from visual data, such as images or videos [54]. In the context of rehabilitation, CV plays a significant role in developing intelligent systems that can assist clinicians, therapists, and patients in various aspects of rehabilitation therapy. By analyzing visual information, CV systems can provide real-time feedback, objective assessments, and personalized guidance during rehabilitation exercises.

The techniques employed by CV resort to algorithms and models to process and interpret visual data, enabling them to perform tasks such as pose estimation, motion analysis, activity recognition, and gesture recognition [15]. These tasks can be interesting for rehabilitation purposes, as they allow us to estimate human pose and evaluate movement quality [15].

2.1.2 Relevant Tasks

This section will primarily focus on two critical CV tasks that play a vital role in rehabilitation: pose estimation and motion quality assessment. These tasks form the fundamental functions of any vision-based system designed for rehabilitation purposes.

2.1.2.1 Pose Estimation

Pose estimation refers to the process of determining the spatial positions and orientations of human joints or body parts from visual input. It typically involves analyzing visual data, such as images or video sequences, and inferring the spatial positions of human joints or body parts [17]. It often relies on the detection of key body landmarks, such as joints or skeletal points, and their spatial relationships. This can be achieved through various ways, including classical approaches, such as pictorial structures, a model that represents the body through its parts and their connections, arranged in a deformable configuration, or DL-based approaches, which employ NNs to directly predict joint positions [17]. In Figure 2.1, we can see an example of Pose Estimation.

Figure 2.1: Example of Pose Estimation

In the context of rehabilitation, accurate pose estimation plays a crucial role in tracking and analyzing patients' movements during exercise or therapy sessions. It provides quantitative measurements of joint angles, limb positions, and body posture, which are essential for assessing and monitoring the progress of rehabilitation programs.

2.1.2.2 Motion Quality Assessment

Motion quality assessment involves evaluating the quality and correctness of movement patterns exhibited by individuals, in our case during rehabilitation exercises. This can include analyzing joint trajectories, comparing against reference movement patterns, and detecting deviations. These assessments are normally done via distance-based methods, statistical analysis, or DL, enabling the identification of incorrect movement patterns and providing valuable feedback [55].

By assessing motion quality, rehabilitation professionals can identify areas of improvement and prevent potential injuries or complications. Furthermore, motion quality assessment provides objective measures for monitoring progress, setting benchmarks, and comparing outcomes across different treatment modalities [55].

2.2 Neural Networks and Deep Learning for Computer Vision

2.2.1 Neural Networks

According to Goodfellow et al. [22], NN models are inspired by the structure and functioning of the human brain. They consist of interconnected nodes, called artificial neurons or units, together in layers. Each neuron receives input signals, applies a mathematical transformation to those inputs, and produces an output. They are called networks because they are organized as a chain of layers that successively take the previous layers' output as input.

They are typically organized as an input layer, followed by different layers with different purposes, followed by a fully-connected layer that produces the output of the network. They are trained on a large amount of data to learn patterns, relationships, and representations within the data [6].

In the context of CV, NNs have been instrumental in advancing the field by enabling the analysis and interpretation of visual data [57]. Two key types of NNs are Convolutional NNs (CNNs) and Recurrent NNs (RNNs).

2.2.1.1 Convolutional Neural Networks

CNNs [32] are particularly well-suited for processing and understanding images due to their ability to exploit the spatial structure of visual data. A CNN's main building block is the convolutional block, besides the input and fully-connected layers. The convolutional block includes three layers: a first layer that actually performs the convolutions, producing a set of linear activations; a second layer that runs each linear activation through a non-linear activation function; and a final layer that pools from a local region to down-sample the output [22].

The convolution is an operation where a small matrix called a kernel is systematically applied to an input, sliding over the input in small steps, performing element-wise multiplication with the local region of the image, and then summing up the results. This process is repeated across the entire image, creating a new output known as the feature map.

The motivation behind using CNNs in CV is that they can capture spatial hierarchies and relationships within the data, allowing the model to learn and recognize patterns such as edges, textures, and more complex structures.

2.2.1.2 Recurrent Neural Networks

RNNs [49] are designed to handle sequential data [22], which is often encountered in CV tasks such as video analysis and action recognition. Unlike feedforward NNs like CNNs, which process inputs independently and in a consistent direction, RNNs have connections that form directed cycles, maintaining an internal state that allows them to capture temporal dependencies and model sequential patterns. Each step in an RNN receives an input and produces an output while also updating its internal state. This recurrent connection allows RNNs to consider the context and history of previous inputs, making them well-suited for tasks involving temporal dynamics [22].

In CV, RNNs are commonly used to analyze sequential data such as video frames, time-series data, or sequences of images. They can capture motion patterns, recognize gestures, and generate sequential predictions [22].

2.2.1.3 Deep Learning

DL is a subfield of machine learning that focuses on training artificial NNs with multiple layers to learn hierarchical representations of data [22]. While traditional machine learning approaches often require handcrafted features, DL models can automatically learn features directly from raw input data. This allows for the extraction of intricate patterns from complex data.

DL has significantly impacted the field of CV by enabling the development of highly effective vision systems [57]. It has demonstrated several advantages in handling complex vision tasks and offers automatic feature learning and generalization abilities.

2.3 Summary

The integration of DL, particularly in the form of NNs such as CNNs and RNNs, within the context of CV systems for accessible rehabilitation holds tremendous potential. Each type of network has its advantages when contributing to the application of CV to rehabilitation. CNNs can extract meaningful features from visual data, enabling accurate analysis of patient movements and actions. RNNs, on the other hand, provide insights into the temporal dynamics of patient movements, enabling more comprehensive evaluations and interventions. Integrating CNNs, RNNs, or both in rehabilitation systems has the potential to enhance the analysis and improvement of rehabilitation outcomes.

Chapter 3

State of the art

The advent of intelligent technology has sparked great interest in the field of rehabilitation, as its integration could enable new systems that have the potential to improve the efficiency of rehabilitation, as they could help ensure exercises are performed correctly. They can also provide personalized assessments and progress tracking, among other benefits. Furthermore, this integration could increase the accessibility of rehabilitation, as these intelligent technologies have the potential to be autonomous and not require continuous professional supervision.

This chapter reviews the state of the art on the application of CV to rehabilitation. Firstly, we present a historical overview of vision-based motion tracking systems for rehabilitation. After that, there is a literature review on the specific topic of motion tracking systems for intelligent rehabilitation.

3.1 Historical Overview

According to the review by Colyer et al. [14], the first attempts to build vision-based motion tracking systems for rehabilitation purposes were based on manual digitization, which consisted of manually localizing points of interest, most typically joints, in each sequential image. This process is very time-consuming, demanding significant effort from experts to identify and track the relevant anatomical landmarks accurately, and can be subject to human errors, introducing potential inaccuracies in the analysis of exercises. The inherent drawbacks served as motivation to search for automated solutions. Automatic systems aim to automate the process of point localization and tracking, alleviating the burden of manual intervention and minimizing human errors. They also offer the advantage of reducing the time required for exercise analysis.

The first automated approach was automatic marker-based motion tracking systems [14]. As the name suggests, these systems relied on the placement of passive markers on the subject's body, in strategic places such as joints, to make it easier to detect the position of these points of interest. This can be used to deduce the pose of the subject. The main advantages of these systems are automation and all that comes with it, such as eliminating the need for human interference and consuming less time, and the accuracy they present. Despite that, the requirement for physical markers causes many drawbacks. Markers take time to place and prepare; they are difficult to place in their correct positions, and that placement can shift from day to day; as they are not completely fixed to a joint, their position can differ from the joint's position; they limit and constrain the subject's movements. These limitations were the driving force for the next step: exploring markerless solutions.

Nowadays, research has focused intensively on markerless motion tracking systems [25], as they present numerous advantages, namely not requiring preparation of the subject, not being invasive or constraining the subject, and being much more accessible [14]. This direction allows for more intelligent methods to be used, namely CV methods, that can detect pose and assess motion quality comparatively. The next sections detail the current state of research on markerless motion tracking systems for intelligent rehabilitation.

3.2 Accessible Motion Tracking Systems for Intelligent Rehabilitation

In recent years, as this is an open research area, some motion tracking systems have been developed for intelligent rehabilitation for research purposes.

Following the structure proposed by [15], in the domain of motion tracking systems for rehabilitation, systems are typically assembled in steps following a given flow. This flow tends to start with *Data Acquisition*, meaning the capture of a raw video of an exercise being performed by a given sensor. Then, *Feature Engineering* takes place, that is, meaningful features are extracted, and after that, these features are encoded, which is done by changing the way the features are represented. In this domain, the representation tends to be a time sequence of a group of joint positions or angles or other kinematic parameters. After that, there is a *Comparison & Assessment* step, where this sequence is compared to another, typically a reference or a desired sequence, in some way to derive an assessment of the quality of the performed exercise that will be the feedback the user will receive.

This can be represented by the diagram in Figure 3.1.

Each step will be detailed in the following sections. A commercial solution will also be presented.

Our analysis is based on 13 identified scientific articles focusing on CV for rehabilitation, which are listed in Table 3.1. The table categorizes the articles based on choices made in terms

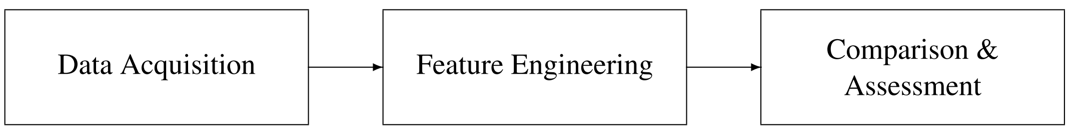

Figure 3.1: Typical Structure

of *Data Acquisition* methods, *Feature Engineering* techniques, and *Comparison & Assessment* methods.

The articles that do not perform any of these steps or fail to mention it are under “Not specified” and if those steps are performed but fall out of the scope of this review (i.e., the last step is not focused on rehabilitation), the cell will be marked “N/A”

By examining the distribution of articles in Table 3.1, we can see that most researchers use *Microsoft Kinect* for data acquisition, although there is significant use of smartphone cameras as well. Regarding feature engineering, most articles that mention using *Microsoft Kinect* for data acquisition adopted the feature engineering tools present in the Kinect Software Development Kit (SDK). Section 3.2.1 provides more information on this matter. Outside of that, the research papers tend to use other feature engineering tools such as OpenPose [8] or BlazePose [5]. The choices made in comparison and assessment methods show some diversity, although there is a clear tendency to resort to NNs.

This highlights the exploratory nature of the research, with different studies adopting varying approaches to address the specific challenges and requirements of intelligent rehabilitation systems.

The table provides an overview of existing motion tracking systems and their characteristics, demonstrating the wide range of approaches and techniques employed in CV for rehabilitation

Article	Data Acquisition	Feature Engineering	Comparison & Assessment
[12]	Microsoft Kinect	OpenNI	DTW
[53]	Microsoft Kinect	Kinect SDK	ANFIS for performance and speed components and Fuzzy Logic to combine those into a score
[10]	Microsoft Kinect	Kinect SDK	HSMM
[16]	Microsoft Kinect	Kinect SDK	N/A
[65]	Microsoft Kinect	Kinect SDK	N/A
[13]	RGB Camera	N/A	Custom scoring function and Fuzzy Logic to obtain an overall score
[21]	RGB Camera	OpenPose	Modular NN
[33]	RGB Camera	PoseNet	Not Specified
[64]	RGB Camera	BlazePose	N/A
[1]	RGB Camera	OpenFace	N/A
[20]	RGB Camera	OpenPose	N/A
[34]	RGB Camera	Convolutional Pose Machines	N/A
[35]	N/A	N/A	Deep NN and GMM log-likelihood performance metric

Table 3.1: Scientific Articles on Motion Tracking Systems for Rehabilitation

research.

3.2.1 Data acquisition

From our analysis of existing literature on accessible technology, we discovered that there are two main options for data acquisition: *Microsoft Kinect* and *RGB Cameras*.

Microsoft Kinect was widely adopted for CV research in rehabilitation. Numerous reviewed articles in this field have relied on *Microsoft Kinect* as their primary data collection tool [12, 53, 10, 16, 65]. After its introduction in 2010, *Microsoft Kinect* rapidly gained popularity and was extensively incorporated into research related to CV in rehabilitation. One of the key factors contributing to its widespread adoption is its affordability, as it was a cost-effective solution for capturing RGB-D (color and depth) images. Unlike traditional RGB cameras, besides color and brightness information, *Microsoft Kinect* provides depth information for each pixel, which denotes the distance of that point from the sensor, according to [14]. *Microsoft Kinect*'s accuracy was reviewed and found to be superior to RGB-only systems [44]. The study conducted by Mousavi Hondori and Kademi deemed *Microsoft Kinect* acceptable for rehabilitation purposes, affirming its suitability as a reliable sensor in this domain. In addition to its low cost and effective RGB-D imaging capabilities, it is important to highlight that *Microsoft Kinect* also included an SDK. This SDK provided researchers with a comprehensive set of tools and libraries that facilitated the development of CV applications tailored explicitly for *Microsoft Kinect*. This made *Microsoft Kinect* much more valuable for researchers.

However, it is important to mention that *Microsoft Kinect* has drawbacks worth considering. These drawbacks include, but are not limited to, the following: first, Microsoft has discontinued production of the *Microsoft Kinect* [41]. This means that researchers are forced to switch to another device or go in a different direction. Additionally, although *Microsoft Kinect* was a low-cost solution compared to some alternative motion capture systems, it, as any other depth-sensor-based system, still represents an additional equipment requirement. Its utilization requires a dedicated *Microsoft Kinect* sensor, which moves away from the original goal of obtaining an as-accessible-as-possible solution.

The main alternative to *Microsoft Kinect* in the studied articles is the *RGB Camera*, which was the approach adopted by most of the reviewed articles [13, 21, 33, 64, 1, 20, 34]. In this set, we include standard, computer, and smartphone cameras, with the latter being the most accessible option.

Smartphones have become ubiquitous and readily available for a significant part of the population. Leveraging smartphones' built-in cameras offers a more accessible approach to video capture, as the need for any extra equipment would not limit patients.

While the main limitation of an *RGB Camera* is its lack of precision compared to a system with a depth sensor, smartphone cameras have nonetheless been deemed suitable for clinical use [43]. In fact, Lam et al. [31] have concluded that smartphones will play a crucial role in the future of the application of CV for rehabilitation, contributing to accessibility.

Table 3.2 summarizes the Data Acquisition component for the reviewed articles.

References	Data Acquisition Method
[12], [53], [10], [16], [65]	Microsoft Kinect
[13], [21], [33], [64], [1], [20], [34]	RGB camera

Table 3.2: Reviewed Articles - Data Acquisition Methods

3.2.2 Feature Engineering

Feature engineering is the second step in the flow of a motion tracking system. It can encompass many smaller parts, including, but not limited to, feature extraction and feature encoding.

Feature encoding is almost never detailed in the studied articles, as features are typically encoded as a time sequence.

Thus, we will focus on feature extraction. Here, the CV task to be addressed is pose estimation. Typically, information deemed relevant is joint positions and angles. Obtaining this information involves tracking and extracting the spatial coordinates of key joints or measuring the angles between body segments during exercises. The extracted features are temporal sequences of the mentioned key points based on their evolution over time. Additionally, various measures can be incorporated, such as the range of an angle or the velocity of a joint, to provide more comprehensive information about movement dynamics.

Our study of the existing literature found that authors vary in their choice of tools for feature extraction. These choices include *OpenPose* [8], *BlazePose* [5], *OpenNI* [19], *FaceMesh*, *OpenFace*, *PoseNet* [28] and *Convolutional Pose Machines* [58].

When *Microsoft Kinect* is used for data acquisition, the provided SDK offers tools that allow developers to access pre-computed skeletal joint positions, orientations, and other relevant parameters. This means there is no need for explicit feature extraction. Therefore, almost all the studied articles that resort to *Microsoft Kinect* do not perform explicit feature extraction, making use of *Microsoft Kinect*'s SDK. The one exception is the work of Chen et al. [12], as they employ a custom process involving a series of image transformations using *OpenNI* [19] to extract spatial information, transforming this visual data into seven key points that they deem required for their rehabilitation system.

When it comes to raw images, feature extraction has to be performed. In the case of rehabilitation, as mentioned before, the task to be addressed is pose estimation.

OpenPose [8] is the most popular choice, being used by Francisco and Rodrigues [21] and Ferrer-Mallol et al. [20], but it has the drawback of being computationally expensive, at least in comparison to other options, as will be seen in the next paragraph.

Another pose estimation library, *BlazePose* [5], was employed by Yang et al. [64], offering similar results to *OpenPose* [8] but with significantly improved computational efficiency, making it apt to use with a smartphone. *BlazePose* [5] is favored for its rapid execution and is integrated into Google's MediaPipe pose detector.

Abbas et al. [1] used *OpenFace* [3], a mobile-oriented facial behavior analysis toolkit. OpenFace focuses on head pose and facial expressions, making it valuable for rehabilitation assessments but limited to specific scenarios.

For real-time human pose estimation, Leechaikul and Charoenseang [33] employed *PoseNet* [28], a lightweight deep learning model. While it sacrifices some accuracy due to its lightweight nature, it serves the purpose efficiently.

Li et al. [34] utilized *Convolutional Pose Machines* (CPMs) [58] in their work. CPMs leverage a CNN architecture with an iterative approach, generating heatmaps at each stage to estimate skeleton key points.

The fact that most reviewed articles used different tools showcases the diversity of the feature engineering domain, as can be seen in Table 3.3.

References	Feature Engineering Method
[12]	OpenNI
[21], [20]	OpenPose
[64]	BlazePose
[1]	OpenFace
[33]	PoseNet
[34]	Convolutional Pose Machines

Table 3.3: Feature Engineering Methods in Reviewed Articles

3.2.3 Comparison and Assessment

Once the relevant features, such as joint positions or angles, have been extracted and represented, they need to be compared and evaluated to assess the quality and effectiveness of the performed exercises. Feature comparison involves the analysis of the extracted features to determine how closely they align with reference movement patterns. Various techniques can be employed for feature comparison, including distance-based, model-less metrics, statistical-model-based metrics, or even Deep Learning. These approaches aim to quantify the similarity between the observed movement patterns and the desired ones.

Reviewing the existing literature on this topic, we can conclude that this portion of the process is the less standardized one, with most articles proposing their own version of a solution by combining several methods. That being said, one proposed option is a direct distance-based technique, Dynamic Time-Warping (DTW). However, most methods are model-based, be it via DL or by Hidden Markov Models (HMMs). Fuzzy logic is also present in more than one article.

Chen et al. [12] developed a vision-based rehabilitation system with a primary focus on action identification, discerning if a subject is performing a rehabilitation exercise, and identifying the specific exercise. They employed DTW to measure the similarity between sequences of varying lengths, enabling comparison of different exercise paces. Their system achieved impressive results, boasting a 98.1% accuracy rate for action identification.

Su et al. [53] employed DTW for comparison and implemented an Adaptive Neuro-Fuzzy Inference System (ANFIS) for evaluation. ANFIS combines NN and fuzzy logic principles. Their evaluation module includes a trajectory evaluator based on a Sugeno-type ANFIS, a speed evaluator, and an overall performance evaluator based on a Mamdani fuzzy inference model. The system matched physicians' scores at an 80.1% rate, demonstrating its effectiveness.

Capecci et al. [10] devised an innovative approach to evaluate rehabilitation exercises using a Hidden Semi-Markov Model (HSMM). Unlike traditional HMMs, HSMM models state durations as distributions. Collaborating with clinicians, they trained the HSMM with data from the seven top-rated subjects. The HSMM calculates a comprehensive score based on observation likelihood. Through hyperparameter tuning, they optimized the HSMM to match clinician ratings. This resulted in a significant correlation for two of the five exercises. Compared to DTW, the HSMM outperformed DTW in correlating with clinician scores for four of the five exercises.

Ciabattoni et al. [13] collaborated with physiotherapists to establish target values, incorporating tolerances for five exercises. They designed a score function based on these targets and measured values for specific joint angles. The overall exercise score is computed as the average or maximum of individual scores. They used a Takagi-Sugeno Fuzzy Inference System to determine the subject's global score, offering flexibility in weighting exercise targets according to physiotherapist recommendations.

Francisco and Rodrigues [21] utilize a Modular NN for exercise assessment. They begin by extracting and storing joint angles, focusing on four angles. After that, their system is divided into their detection module, which identifies what exercise is being performed, and their measure module, which employs an NN to assess exercise correctness. The initial network architecture was straightforward and comprised three layers. Multiple configurations were tested to obtain the best architecture and evaluated using accuracy and the area under the Receiver Operating Characteristic (ROC) curve. Ultimately, the most effective architecture ended up featuring eight hidden layers and achieving 94.54% accuracy and an area under the ROC curve ranging from 0.8 to 1, depending on the exercise.

Liao et al. [35] employ a Gaussian Mixture Model (GMM) log-likelihood performance metric for rehabilitation, leveraging its ability to effectively capture the inherent variability in human movements. This metric is incorporated into a deep-learning model comprised of a complex deep NN that includes convolutional layers, recurrent layers, and temporal pyramids. It outputs movement quality scores for input exercises. In their evaluation, they compared the GMM log-likelihood metric to Euclidean distance, Mahalanobis distance, and DTW distance, focusing on its capacity to distinguish correct from incorrect movements. The results showcased the superiority of the proposed metric, although DTW distance and Euclidean distance displayed noteworthy results. Additionally, they assessed their deep-learning model against other NN models, using absolute deviation as the performance metric. The model demonstrated overall superior performance, although specific models outperformed it in certain exercises within the dataset.

Table 3.4 showcases the diversity in Comparison & Assessment methods.

Reference	Comparison and Assessment Method
[12]	DTW
[53]	ANFIS for their performance and speed components and Fuzzy Logic to combine those into a score
[10]	HSMM
[13]	Custom scoring function and Fuzzy Logic to obtain overall score
[21]	Modular NN
[35]	Deep NN and GMM log-likelihood performance metric

Table 3.4: Comparison and Assessment Methods in Reviewed Articles

3.3 Commercial Systems

In addition to research efforts, there are also commercial solutions available in the field of CV for rehabilitation. One such solution is Exer Health [18], a patient mobile app designed to enhance the rehabilitation process. Exer Health aims to keep patients engaged and gather critical health assessment data while they perform exercises at home. The app claims to measure range of motion, counts repetitions, recommends form adjustments, and provides real-time feedback to patients. This feedback helps patients adhere to their recovery protocols, and the data collected can be used by professionals to evaluate progress throughout the rehabilitation journey. Moreover, Exer Health offers professionals an intuitive mobile app that facilitates the creation of high-touch, closed-loop recovery protocols. This gives patients a richer experience while providers focus on delivering optimal care.

Regarding technology, Exer Health claims to employ a proprietary motion-AI platform that powers all of its digital health software. The platform is said to run "on the edge," utilizing common laptops, phones, and tablets. This approach ensures accessibility and ease of use for patients and professionals, allowing seamless technology integration into rehabilitation. Although they do not give any details, they mention the use of NNs, machine learning, and CV.

Figure 3.2 shows an example of how the Exer Health application works.

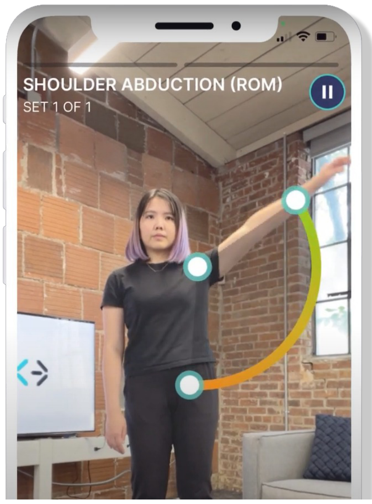

Figure 3.2: Exer Health App. Source: Exex Health [18]

The main problem with Exer's solution is that they do not reveal much about it, with their explanations being very generic when it comes to the technology used and how they used it. Our work differs from theirs in that sense, as our objective is to make this research public and available for anyone to understand.

3.4 Available Datasets for Modeling Computer Vision Systems

The availability of high-quality and diverse datasets is very important in training CV models for rehabilitation applications. Whereas in other motion modeling applications based on CV, there is a wide range of large public datasets, in the case of rehabilitation data, authors tend to collect their own data, obtaining relatively small datasets [15]. We compiled the most relevant datasets found while researching in Table 3.5.

Dataset	Number of Individuals	Number of Exercises
IntelliRehabDS [42]	29	9
UI-PRMD [56]	10	10
KIMORE [9]	78	5

Table 3.5: Relevant datasets for CV-based rehabilitation

However, these datasets all mention that their data is captured by an RGB-D camera such as *Microsoft Kinect*. This is also true for other datasets found during the literature review, as well as others that mention the use of body-worn sensors. This is a limitation of the current state of this research field.

It is also worth mentioning that in recent years, the emergence of Generative Adversarial Networks (GANs) has provided a promising avenue for expanding the available data in rehabilitation research. GANs can generate synthetic data that closely resembles real rehabilitation exercises, thereby augmenting the existing datasets. This approach has the potential to address the limitations of small-scale datasets and facilitate the development of more robust and generalizable CV models. Li et al. [34] develop GANs to augment the UI-PRMD dataset.

3.5 Takeaways and Conclusion

The previous sections have explored the intersection of CV and intelligent rehabilitation systems, examining the relevant literature and gaining insights into this field. We first gave a historical overview and then dove deeper into Accessible Computer Vision Systems for Intelligent Rehabilitation, focusing on three critical aspects of these systems: data acquisition, feature engineering, and comparison & assessment.

We observed that data acquisition methods have evolved over time. While *Microsoft Kinect* was once a prominent choice, its discontinuation has led to the emergence of smartphone cameras as accessible and reliable alternatives for rehabilitation purposes.

In terms of feature engineering techniques, OpenPose is the most popular choice. Although it is widely used in research that requires pose estimation, it is computationally expensive, so we highlighted an alternative, BlazePose, that offers greater efficiency and speed.

The comparison and assessment of exercises involve diverse techniques, including DTW, fuzzy inference systems, HSMMs, and deep NNs. These approaches aim to quantify the similarity between observed and desired movement patterns. Every studied article mentioned satisfactory results.

What we conclude by comparing the different studies is that DTW seems to be a standard method. It could be interesting to use as an initial method and a term for comparison of different methods, serving as a benchmark. We also found that the trend is the use of DL, as in most fields of computer science. Fuzzy logic seems to be regarded as an effective tool for this task, as well as HMMs.

We also presented a commercial solution, which seems to be promising, but no concrete details are given in terms of the technology used.

Another important aspect mentioned is the available datasets. As detailed before, there are not many large public datasets, which is something researchers have to deal with. Typically, the solution is for authors to collect their own data.

Building on the insights from this review, our work will take the discontinuation of *Microsoft Kinect* into consideration and explore the potential benefits of smartphone cameras. DTW represents a widely accepted approach and will be employed as a preliminary method, serving as a reference for comparison. DL is emerging as a dominant trend, and future work should aim to maximize its potential. This could involve the development of sophisticated DL models and exploring their integration with fuzzy logic or HMMs.

Chapter 4

Problem Statement

This chapter defines the problem to be addressed by this thesis. The first section recaps our context and motivation. Then, there is a section that clearly identifies the problem, addressing the limitations and challenges associated with current rehabilitation methods. The next section then presents the research questions we must answer. In the end, there is a brief conclusion for the chapter.

4.1 Recap

As mentioned in Section 1.1, traditional rehabilitation approaches are often time-consuming, expensive, and dependent on the availability of healthcare professionals. This sets an undesirable limit on the number of sessions a patient can attend and may lead to a decrease in the speed of their recovery progress [39].

The initial solution would be to follow at-home rehabilitation methods, with the healthcare professional prescribing exercises that the patient can perform at home. These methods, although convenient, come with their own challenges. Without proper feedback and guidance, there is a risk of performing exercises incorrectly, leading to a slower recovery or, in a worst-case scenario, aggravating injuries. Additionally, it is very common to observe low adherence to prescribed exercises [4].

Recent advances in technology could offer some alternatives. While robotic-assisted rehabilitation offers potential benefits, it is often limited to specific conditions and comes with a high cost. VR systems, although immersive, require additional equipment and the creation of a virtual world, which can be a barrier to widespread adoption. Similarly, many CV-based solutions in rehabilitation are marker-based, relying on specialized equipment or setups, making them less accessible and suitable for at-home use, as shown in Chapter 3.

However, the interest in markerless motion tracking systems has grown, as they hold the potential to significantly improve the recovery process and the accessibility of rehabilitation services. By integrating these systems, patients would be allowed to perform rehabilitation exercises independently, without the need for the availability of a healthcare professional, while receiving correct feedback. This approach eliminates the constraints of in-person sessions and makes rehabilitation more accessible.

From the perspective of the patient, this solution would empower them, greatly increasing their access to rehabilitation and speeding up their recovery process. For healthcare professionals, this could mean they could possibly track patients' progress remotely and see their customers recover independently.

4.2 Problem Statement

Following what is mentioned in the previous section, the main problem in the field of intelligent rehabilitation is accessibility, whether it be in terms of financial costs, geographical limitations or healthcare professional availability. Despite some solutions having been proposed, it has been demonstrated that this problem remains unresolved, leading to limited access to quality rehabilitation services for individuals.

In more practical terms, we want to solve the problem of accessibility in rehabilitation using technology, more specifically Computer Vision. We want to target the patients and give them accessible tools to perform exercises independently. By doing this, we would enhance the availability and effectiveness of rehabilitation services, improve patient outcomes, and revolutionize the broader healthcare ecosystem, as this solution would result in reduced healthcare costs and improved resource allocation.

4.3 Research Questions

There are specific implementation challenges that need to be addressed to develop an effective solution for accessible rehabilitation.

The primary research question is how to assess the quality of the execution of a rehabilitation exercise in an accessible way using CV. This breaks down into other smaller questions:

1. How can CV retrieve information relevant to assessing the quality of rehabilitation from a video?

2. How can that information be used to effectively assess the quality of rehabilitation?

The first question goes into the aforementioned Pose Estimation task, which is explained in Section 2.1.2.1. By accurately estimating human pose, our proposed solution aims to provide information about aspects such as body alignment, joint angles, and movement trajectories. Resolving this research question is vital to building an accurate and effective accessible rehabilitation system.

The second question is focused on Motion Quality Assessment, mentioned in Section 2.1.2.2. Our solution will seek to accurately assess motion quality.

Another very important aspect to prioritize that is common to all research questions is accessibility. When researching, we want to find the most accessible way to answer these questions and always choose the most accessible options for our system.

Additionally, determining the criteria for evaluating exercise performance and establishing meaningful metrics to quantify quality are important considerations.

4.4 Conclusion

In this chapter, we have discussed the problem at hand. The limitations and challenges associated with current rehabilitation methods have highlighted the need for a more accessible approach. By integrating CV, we aim to create a platform that allows patients to perform rehabilitation exercises independently while receiving adequate feedback.

Chapter 5

Implementation

5.1 Introduction

This chapter explores the implementation process of our computer vision-based motion tracking system for accessible rehabilitation, building upon the theoretical principles presented in the previous chapters. We will discuss technology selection and present the data we worked with. Then, the focus will be on the architecture of the system. After that, we will address our system's limitations and briefly summarize our implementation process.

5.2 Architecture

Our system follows the structure presented in Chapter 3, meaning it involves three components: Data Acquisition, Feature Engineering, and Comparison & Assessment. This means we have to make technological and architectural choices for each one of these components. The following subsections will detail the choices made for each of the components.

5.2.1 Data Acquisition: Smartphone Camera

In the Data Acquisition component, we had to choose between *Microsoft Kinect* and the RGB camera, as is explained in Section 3.2.1. Deriving from what is mentioned in the aforementioned Section, we opted for the RGB camera, more specifically, the smartphone camera, due to the ubiquity of the smartphone, aligning with the focus on accessibility. Furthermore, it should allow for a more straightforward and user-friendly approach to data collection.

5.2.2 Feature Engineering: MediaPipe's Pose Landmarker (BlazePose)

Regarding Feature Engineering, the range of choices was much broader, as detailed in section 3.2.2. The most popular choice, OpenPose, had the drawback of being computationally expensive, which could pose problems given the choice of the smartphone, a less computationally powerful device. One of the alternatives was BlazePose, which presented results similar to OpenPose but was much less computationally expensive, making it suitable for a smartphone. One framework that implements the BlazePose algorithm is the MediaPipe Pose Landmarker.

MediaPipe [23, 38] is an open-source framework created by Google that provides a toolkit for the development of Machine Learning applications, mainly focused on helping people design and implement solutions for vision, audio, and text-based machine learning tasks. In terms of audio and text, it provides developers with tools suitable for various classification, detection, and embedding tasks. When it comes to vision-based tasks, it offers tools to perform Object Detection, Image Classification, Hand Landmark, Hand Gesture Recognition, Image Segmentation, Interactive Segmentation, Face Detection, Face Landmark Detection, and the one we're interested in, Pose Landmark Detection.

The MediaPipe Pose Landmarker [24] is a component within the broader MediaPipe framework designed for human pose estimation. This specific module is focused on identifying and tracking key points, or what they call pose landmarks, on a person's body, enabling applications to access an estimate of the human pose. It receives as input still images, decoded video frames, or live video feed and outputs pose landmarks in normalized image coordinates or pose landmarks in world coordinates. It can also render an overlay of the landmarks and their connections over the original input. It is also important to mention that it can detect multiple poses, with the maximum number of poses to detect being defined by the developer. The developer can also set values for the minimum confidence scores to consider pose detection and pose tracking successful, as well as for the minimum confidence score of pose presence score in the pose landmark detection. This component of MediaPipe includes two models: a pose detection model, including just a few key landmarks, and a pose landmarker model, which completely maps the pose, outputting 33 key landmarks. Figure 5.1 shows the position of the 33 landmarks in the human body, and Table 5.1 shows the index of each landmark.

Index	**Body Part**
0	Nose
1	Left Eye (Inner)
2	Left Eye
3	Left Eye (Outer)
4	Right Eye (Inner)
5	Right Eye
6	Right Eye (Outer)
7	Left Ear
8	Right Ear
9	Mouth (Left)
10	Mouth (Right)

(a) Part 1

Index	**Body Part**
11	Left Shoulder
12	Right Shoulder
13	Left Elbow
14	Right Elbow
15	Left Wrist
16	Right Wrist
17	Left Pinky
18	Right Pinky
19	Left Index
20	Right Index
21	Left Thumb

(b) Part 2

Index	**Body Part**
22	Right Thumb
23	Left Hip
24	Right Hip
25	Left Knee
26	Right Knee
27	Left Ankle
28	Right Ankle
29	Left Heel
30	Right Heel
31	Left Foot Index
32	Right Foot Index

(c) Part 3

Table 5.1: MediaPipe Landmarks. Source: Google Mediapipe [24]

The output of the MediaPipe Pose Landmarker contains X, Y, and Z coordinates for each landmark and a visibility factor, representing how likely it is that the landmark is visible. As men-

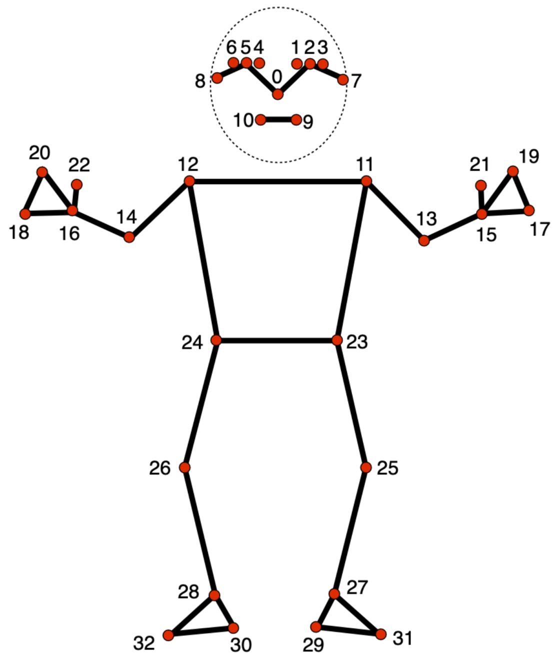

Figure 5.1: MediaPipe Pose Landmarker Model. Source: Google Mediapipe [24]

tioned before, the MediaPipe Pose Landmarker can output pose landmarks in either normalized image coordinates or world coordinates. If the output is normalized image coordinates, X and Y are normalized between 0 and 1 in relation to the image's width and height. The Z coordinate is the landmark depth, with the hips as the origin. It has the same magnitude as x, and the smaller the value is, the closer it is to the camera. If the output is world coordinates, X, Y, and Z are Real-world 3-dimensional coordinates in meters, also with the hips as the origin.

An example of the output of a landmark would be:

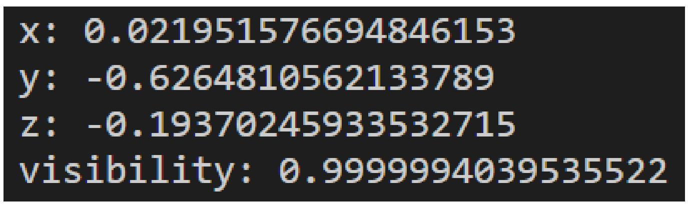

Figure 5.2: Example of Landmark output

It is also important to mention that this module implements a variant of BlazePose [5] that utilizes GHUM [63]. It also uses a CNN similar to MobileNetV2 [52].

BlazePose is a lightweight CNN architecture designed for real-time human pose estimation on mobile devices. It employs a pose detector and a pose tracker. For the pose detector, it is assumed the head of the person is always visible, and a face detector is used to identify the face and predict other parameters, such as the person's orientation. The pose tracker consists of a NN that predicts the location of 33 keypoints of the human body. This NN follows an encoder-decoder architecture that is used to create heatmaps and is followed by another encoder for regression. In order to evaluate the model, it was compared to OpenPose [8] on two datasets and obtained very similar results. The main advantage of BlazePose is that it is 25–75 times faster when running on a single mid-tier phone CPU than OpenPose running on a 20 core desktop CPU.

This variant of BlazePose uses GHUM, a pipeline designed for accurate 3D human shape modeling. GHUM is trained within a Deep Learning framework, having full 3D body scans as input, and is constructed using an end-to-end learning pipeline that enables the training of statistical models based on various body-related parameters. The three main components of the pipeline are the Variational Body Shape Autoencoder for body shape estimation, the Variational Facial Expression Autoencoder to detect facial expressions, and the Skinning Model, which attempts to represent skin and realistically deform the 3D model to match the movement of the underlying skeleton. GHUM's pose estimation is performed based on body shape estimates and skinning parameters.

The CNN used is similar to MobileNetV2, an NN architecture designed for object detection on mobile devices. The main goal is to improve computational efficiency while maintaining accuracy. With that in mind, it introduces a new layer module, the inverted residual with linear bottleneck. It presents very competitive results at a much lower computational cost.

5.2.3 Comparison & Assessment: Dynamic Time Warping

To perform the Comparison and Assessment part of our system, we opted for DTW.

DTW is an algorithm used to assess how similar two time series are by measuring the distance between those two series. Let's use as an example two time series A and B. The simplest way to measure the distance between two time series would be to use Euclidean distance, which is computed as the square root of the sum of the squares of the differences between A_i and B_i, where A_i represents the value of time series A at index i and B_i represents the value of time series B at the same index. Equation 5.1 describes Euclidean distance.

$$EuclideanDist = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{5.1}$$

The problem with this measure is that it does not account for shifts along the time axis. For example, if two time series are equal but one of them is shifted in time, Euclidean distance would indicate a difference between both series.

For example, let's consider the following sequences:

1. $[0,1,2,3,0,0]$

2. $[0,0,1,2,3,0]$

We can understand that the sequences represent the same evolution, but sequence 2 starts later than sequence 1. If we calculate the Euclidean distance between these series, we would obtain a result of $\sqrt{12} \approx 3.46$, when the desirable result would indicate a smaller difference.

With the aim of solving this problem, DTW was developed [50]. The idea behind DTW is the computation of a warp path that warps points of the two series to one another in order to obtain a more intuitive relationship between the series. Warping two points is creating a correspondence between them. This path includes every index of both time series, meaning that every point in one sequence is warped to a point in the other sequence. The goal is to compute the optimal warp path, meaning the one with the minimum distance. To do so, we must first compute a two-dimensional distance matrix, D. Given the same time series A and B, D is $|A|$ by $|B|$, with each axis representing one of the time series. The process of computing the matrix is done dynamically, starting with $D(0,0)$ and finishing with $D(|A|,|B|)$ The value of each cell in D is computed by the following formula:

$$D(i,j) = Dist(A_i, B_j) + \min[D(i-1,j), D(i,j-1), D(i-1,j-1)] \tag{5.2}$$

In every calculation, we add the distance between the two points in each axis at the current index, $Dist(A_i, B_j)$, to the minimum value of the three cells that precede the current one. This is because we know that the optimal path must pass through one of those cells and that one of them contains the minimum possible distance, so we add the current distance to the smallest of those three values. $Dist(A_i, B_j)$ can be calculated by various distance measures, such as Euclidean distance or Manhattan distance. A comparative study between different distance measures for DTW showed that Euclidean distance obtains the best results [30]. After we fill the matrix, the value of the DTW distance between the two sequences is the value at $D(|A|,|B|)$.

For example, let's consider the same sequences 1 and 2. Figure 5.3 shows the filled distance matrix for the calculation of DTW between the sequences.

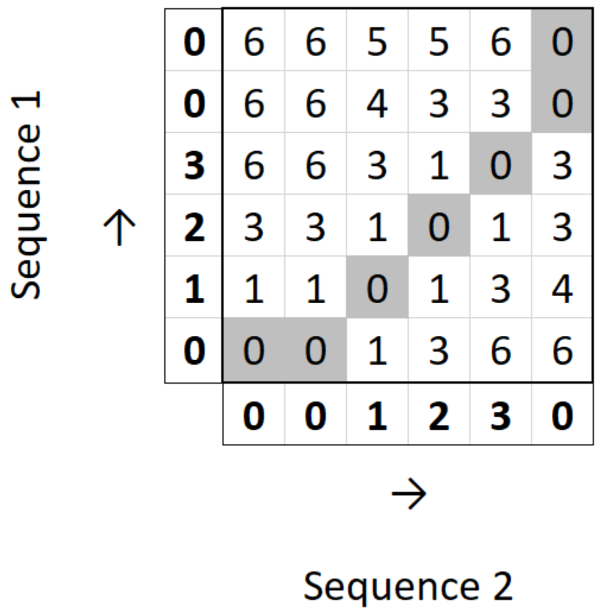

Figure 5.3: Distance matrix for DTW calculation

As we can see, we obtain the optimal path from $D(0,0)$ to $D(|A|,|B|)$ and are left with the value 0 at $D(|A|,|B|)$, which means that the DTW distance between the two sequences is 0.

The choice for DTW stems from the fact that it is simple, being a distance-based metric, but presents impressive results measuring the similarity between sequences of movements.

5.3 Dataset

Collaborating with the Escola Superior de Saúde of the Instituto Politécnico do Porto was instrumental in accessing a pre-existing dataset of exercises. The data that is included was collected with the primary aim of examining rehabilitation exercises. Subsequently, it has been utilized in various works, including the study conducted by Lopes et al. [37] and my own thesis. Fifteen patients, labeled with IDs 1 to 15, were asked to perform two distinct exercises, *Diagonal* and *Rotation*, each repeated three times. Those exercises were recorded with a smartphone camera, and those videos are what constitute our dataset, as well as two reference videos performed by healthcare professionals, one for each video. That is a total of 90 examples, 45 for each exercise, plus the two reference videos. The dataset has been evaluated by an advanced state-of-the-art system, the Qualisys Motion Capture System (QTM) [47], that can evaluate Range of Motion (ROM), being scored from 0-100. However, the reference scores are not attributed to each example in the dataset but to each patient, making a total of 30 scores: one for each of the two exercises performed by the 15 patients. This score represents a movement quality index based on comparing ROM between instances of subjects performing Guided Exercises (GE) and Non-Guided Exercises (NGE) and is calculated by the following equation:

$$MovementQualityIndex(\%) = \left(1 - \frac{ROM_{GE} - ROM_{NGE}}{ROM_{GE}}\right) \times 100 \tag{5.3}$$

Figure 5.4 shows frames of a video from the dataset.

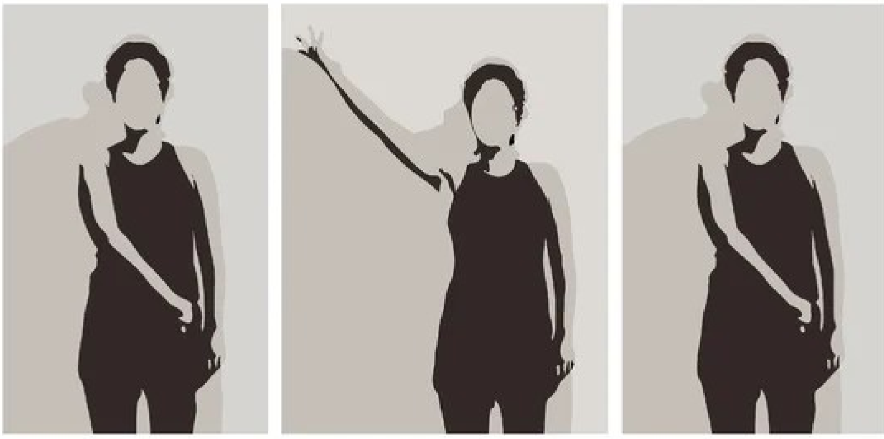

Figure 5.4: Frames of the *Diagonal* exercise. Source: Pereira et al. [45]

The selected dataset contains outliers, which have been explicitly identified by the researchers responsible for its collection. The presence of these outliers has implications for the evaluation of our system, as we must address and handle them effectively. In Chapter 6, Section 6.2 will provide an in-depth exploration of the strategies employed to deal with these outliers.

5.4 Experimental Setup

All of the processing was performed on a personal computer with the following specifications:

1. **CPU**: Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz
2. **RAM**: 8,00 GB
3. **Operating System**: Windows 11 Pro 64-bit
4. **Python Version**: 3.10.11

5.5 System Development

The development journey started with a basic idea of how the system would be structured, based on the studied literature and detailed in Section 5.2, and a simple division of the development plan. The first part of the development would be to build a system capable of comparing two videos, and the second part would involve generalizing to compare all the videos in the dataset to the reference video for the corresponding exercise and evaluating the system. This evaluation would be made by comparing the results obtained by our system and the correctness scores associated with each video, as mentioned in Section 5.3.

5.5.1 Comparing two videos

In order to obtain a system that compares two videos, we devised the following plan:

1. Start by focusing on a single exercise and identify the relevant joints for that exercise.
2. Process one input video:
   (a) Read the video.
   (b) Extract values of relevant joint positions for each frame.
   (c) Create and save a sequence of these values.
3. Process another input video in the same fashion.
4. Calculate the DTW distance between two sequences.

Thus, we started by selecting one of the two mentioned exercises, *Diagonal* and *Rotation*. There were no substantial differentiating factors between the two regarding what would concern comparing examples of each one, so we opted for *Diagonal*. We then wanted to identify which joints were relevant for the exercise. In order to do that in a clinically sound manner, we made use of our collaboration with the Escola Superior de Saúde. We asked them which joints were relevant for comparing this exercise. They mentioned that they divided joints into three major groups for their comparison: Head, Trunk, and Shoulder. They also mentioned which landmarks of the MediaPipe Pose Landmarker Model should be included in each joint group.

This meant that the relevant joints were organized as follows:

- Head: Landmarks 0 to 10
- Trunk: Landmarks 11, 12, 23, 24
- Shoulder: Landmarks 12, 14, 16

The next step was to process an input video. This involved developing a Python script where videos could be processed, and relevant data could be retrieved, particularly the positions of the relevant landmarks for each frame. In order to process the video, we had to resort to the OpenCV library [7]. OpenCV is an open-source software library designed for computer vision and machine learning applications. It offers a comprehensive set of image and video processing tools, including modules for image manipulation, object detection, feature extraction, and more. It is widely utilized in computer vision applications.

We then made use of the MediaPipe Pose Landmarker to perform pose estimation and obtain the X, Y, and Z coordinates of each landmark present in the relevant joint groups for each frame. At this point, we faced a choice between using MediaPipe Pose Landmarker's `Landmarks` or `WorldLandmarks`. We opted for `WorldLandmarks` because they are relative to the subject's body. `Landmarks` are sensitive to the position of the subject in the image, which can cause problems when comparing two videos. These coordinates were saved after each frame, and we created a sequence of coordinates for each relevant landmark. Each sequence was essentially a list that had as elements smaller lists, one for each frame, that had as elements the coordinates of that landmark at that frame. After processing a video, we were left with a set of lists, one for each landmark, representing the evolution of the position of that landmark over time.

A generic sequence for a landmark would be represented as:

$$[[x_{t=1}, y_{t=1}, z_{t=1}], [x_{t=2}, y_{t=2}, z_{t=2}], \ldots, [x_{t=T}, y_{t=T}, z_{t=T}]] \tag{5.4}$$

In this sequence, each list represents the X, Y, and Z coordinates of the landmark for each frame.

After we had processed two input videos and obtained the sequences that represented the motion in terms of the evolution of the position of the relevant joints, we were ready to compare them and assess how similar those sequences were and, consequently, how similar the exercises being performed were.

To compare the two exercises, we opted to calculate the distance between each relevant landmark's evolution in one video to the corresponding one in the other video and then sum up those distances according to the grouping suggested before, meaning the distances between landmarks in the head group would be summed up to obtain the overall distance measure for the head, as was done for the trunk and shoulder landmark groups. In order to calculate the distance between two sequences, we used DTW.

There are many libraries that implement DTW in Python, as is the case of *dtaidistance*, *TSLearn*, or *FastDTW*. *FastDTW* is the most commonly used one [62], but it is actually an approximation, designed to perform the DTW calculation faster than the original algorithm [51], whereas both *dtaidistance* and *TSLearn* perform the full DTW algorithm.

However, it has been claimed that *FastDTW* might actually not be faster and, because it is an approximation, the trade-off between accuracy and speed might not be worth it [62].

In order to make a choice, we compared the execution of the three algorithms as well as a generic ad hoc DTW implementation. To compare the execution time of each algorithm, we ran each one on the same two videos ten times, registering the average of their execution times.

```
DTW:             29.164266586303711 seconds
FASTDTW:          2.317301750183106 seconds
DTAIDISTANCE:     0.090170621871948 seconds
TSLEARN:          4.478096008300781 seconds
```

Figure 5.5: Comparison of times of execution for the different DTW implementations

As we can see in Figure 5.5, *dtaidistance* had the smallest execution time, so that was the final choice.

This meant we could now compare two videos and obtain distance measures for the three landmark groups, concluding the initial part of our development plan.

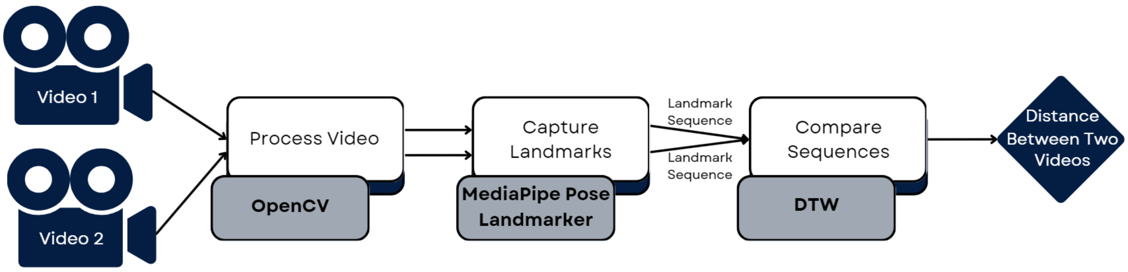

Figure 5.6: Pipeline for comparing 2 videos

Figure 5.6 shows the pipeline followed to develop the part of the system that compares two videos and obtains a distance between them. It takes two videos as input, processes them using OpenCV, captures the landmarks of both using the MediaPipe Pose Landmarker, obtaining the landmark time sequences, which are then compared using DTW, resulting in a measure of the distance between the two videos.

5.5.2 Generalizing for the whole dataset

After we had built a system capable of comparing two videos and obtaining a measure of the difference between the videos, our intention was to generalize and obtain that measure for each exercise in the dataset.

Thus, we wrote a Python script that would do this for us. First, it processed the reference video for one exercise and saved the landmark sequences. Then, it iterated through the dataset, reading each video of that exercise and obtaining the landmark sequences, comparing them, using DTW, to the reference. This gave us a distance score for the three landmark groups in each video, which were then saved in a data frame and, subsequently, in a *.csv* file. This process was repeated for the second exercise, leaving us with two files with the distances recorded between the time sequences of the reference exercise and each of the examples in the dataset. Each file was divided into distances for the Head, Trunk, and Shoulder landmark groups.

These files were then used as input for the evaluation of our system, comparing our results to those obtained by the QTM mentioned in Section 5.3. The evaluation of the system will be discussed in Chapter 6.

ID	Head	Trunk	Shoulder
ID01_1	220.63	50.39	103.13
ID01_2	258.58	55.52	109.38
ID01_3	257.50	55.06	122.51
ID02_1	272.44	37.97	72.85
ID02_2	212.43	32.74	80.76
ID02_3	310.14	47.60	89.19
ID03_1	159.56	32.30	99.23
ID03_2	152.75	29.36	102.72
ID03_3	176.59	31.23	86.32
ID04_1	235.18	54.42	82.61
ID04_2	188.21	47.61	90.15
ID04_3	203.35	60.69	86.92
ID05_1	308.34	70.13	117.46
ID05_2	301.78	67.27	106.83
ID05_3	316.42	65.68	113.23

Table 5.2: First 15 rows of the distances file for the *Diagonal* exercise

Table 5.2 shows a portion of the *.csv* file that contains the distances between the *Diagonal* exercise examples in the dataset and the *Diagonal* reference exercise, divided into the Head, Trunk, and Shoulder landmark groups.

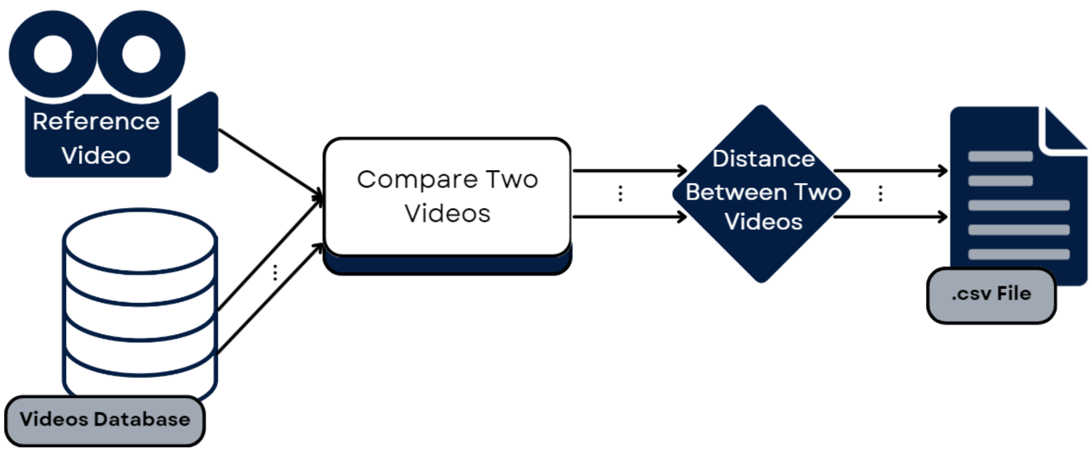

Figure 5.7: Pipeline for comparing the entire dataset to the reference

Figure 5.7 shows the pipeline followed to develop the part of the system that iterates through the dataset, comparing every example video to the reference and obtaining a distance for each one.

5.6 Summary

In this chapter, we detail how our CV-based motion tracking system was implemented. We mentioned the architecture of our system and the choices that were made at each step of the system flow, namely the choice of the smartphone for Data Acquisition, MediaPipe's Pose Landmarker for Feature Engineering, and DTW for Comparison and Assessment. We detailed the dataset made available to us and mentioned the technology used. Lastly, we explored the development journey of our system, explaining how each step is performed.

Chapter 6

System Evaluation and Results

This chapter is focused on the evaluation of our system and the presentation and discussion of the obtained results. The first section explains the concept of data normalization, why it is necessary, and how we opted to perform it in our data. After that comes a section on the details of our system evaluation, where we discuss the metrics used to evaluate and how we handled outliers. Then, there is a section presenting the results and a following section discussing them. After that, there is a section on statistical analysis. Finally, there is a small summary of the chapter.

6.1 Data Normalization

To evaluate the performance of our system, we will compare our results to those obtained by the QTM, mentioned in Section 5.3, which serves as the ground truth for our dataset. However, our results are a distance measure, whereas the dataset is labeled with scores on a scale of 0 to 100, as mentioned in Section 5.3. To ensure a meaningful comparison, we need to make sure that both sets of results are on the same scale, which means we need to transform our distance measures into a 0 to 100 score. One way to achieve this is through data normalization.

Data normalization is a process in which data is scaled to a specific range. It is usually performed as a preprocessing step before the comparison of two sets of data and allows the comparison of data with different ranges. There are many different ways of performing data normalization, including, for example, Z-score normalization or Min-max scaling.

Lima et al. [36] conducted a comprehensive study comparing various normalization options, more specifically for the normalization of time series. They mention that Z-score is by far the most popular option, with many authors not even considering the alternatives. It is particularly well-suited for normally distributed data. However, they conclude that it may not always be the best option and recommend at least exploring one more option before using just Z-score, more particularly Maximum absolute scaling.

The next paragraphs detail each of the three mentioned methods: Z-score normalization, Min-max scaling, and Maximum absolute scaling.

Z-score normalization transforms data to have a mean of 0 and a standard deviation of 1, with each transformed value being calculated with the following equation, where μ is the mean of the data, and σ is the absolute deviation:

$$x' = \frac{x - \mu}{\sigma} \tag{6.1}$$

Min-max scaling scales the data to a specific range, often 0 to 1, but in our case, 0 to 100, based on specified minimum and maximum values. These values are usually the maximum and minimum values of the dataset, but we could select values that correspond to what would be considered, in the original scale, a score of 0 and a score of 100. For example, in our case, the minimum would be 0, as the ideal performance of an exercise corresponds to a distance of 0 when comparing the two time sequences, and the maximum would be the values obtained by comparing two very different videos. One crucial detail to consider is that, in our case, larger distances correspond to lower scores and vice versa, which means that, after scaling is done, we invert the scale by subtracting each score from 100. The transformed values are calculated via the following equation:

$$x' = \frac{x - min}{max - min} \tag{6.2}$$

Maximum absolute scaling works in a very similar way but just considers the maximum absolute value. Each value is obtained with the following equation.

$$x' = \frac{x}{|max|} \tag{6.3}$$

Because all our data is positive and we know the minimum value is 0, Maximum absolute scaling and Min-max scaling have the same behavior. From this point on, we will only refer to Min-max scaling.

Before choosing the normalization technique, we must first analyze our data.

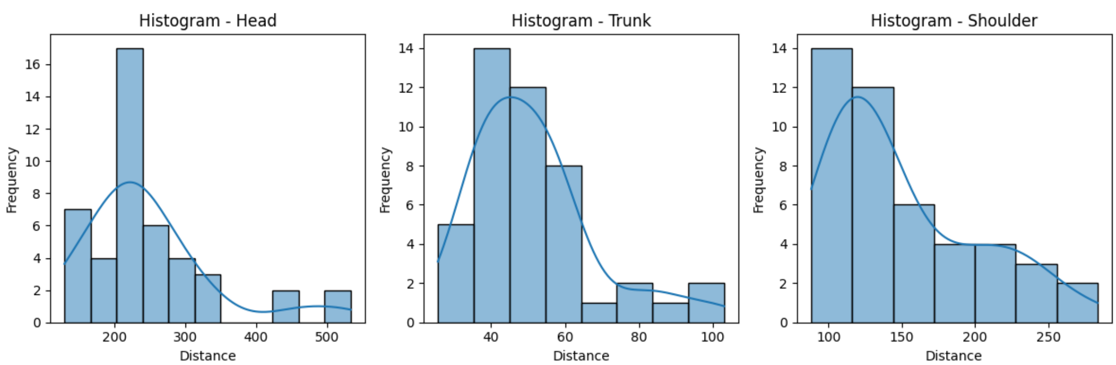

Figure 6.1: Histograms for the distances obtained for the *Rotation* exercise

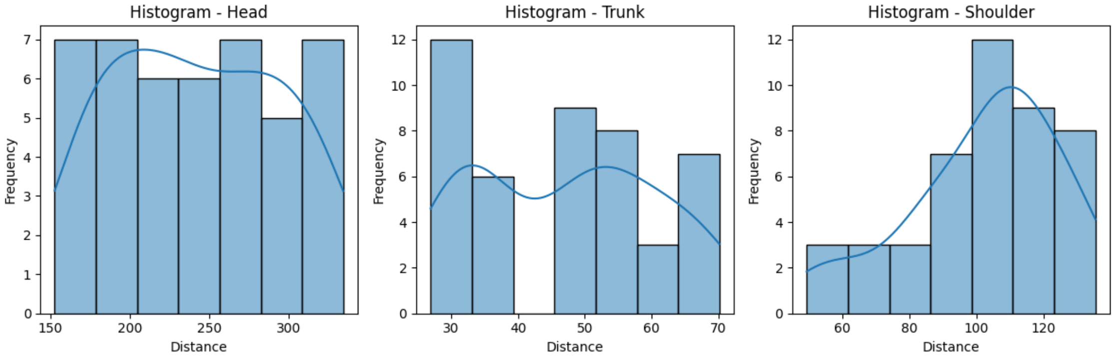

Figure 6.2: Histograms for the distances obtained for the *Diagonal* exercise

Figures 6.1 and 6.2 help us visualize the distribution of our data. They include histograms for the three joint groups for each exercise.

Examining the presented graphs, it becomes evident that the distribution of the data does not have the characteristics of a normal distribution, as there are clear deviations from the expected bell-shaped curve. This tells us that the data does not conform to a normal distribution and that Z-score normalization would not necessarily be suitable for our data.

However, due to its immense popularity, and given most authors will not even consider not using Z-score, we decided to compare the effectiveness of these normalization techniques, applying both Z-score and Min-max scaling to our data.

Finally, we calculate the average scores for three exercises per ID, providing an overall performance score for each exercise. This aggregated score is what we compare to the QTM score. An example of the aggregated scores can be seen in Table 6.1.

ID	**Head**	**Trunk**	**Shoulder**
ID01	60.08	56.45	77.31
ID02	55.95	61.03	76.73
ID03	80.08	72.93	76.14
ID04	65.33	58.81	76.47
ID05	63.32	61.25	80.64
ID06	72.1	73.42	79.74
ID07	70.18	63.21	67.68
ID08	73.92	69.19	65.31
ID09	69.99	66.98	80.01
ID10	57.75	59.85	78.93
ID11	70.94	74.62	78.15
ID12	29.1	29.92	54.14
ID13	65.54	58.5	70.37
ID14	78.75	72.59	82.31
ID15	62.81	48.89	59.19

Table 6.1: Scores file for the *Rotation* exercise using Min-max scaling

6.2 System Evaluation

The evaluation process is very important to assess the performance and practical utility of the system. A thorough evaluation can not only validate a system but also identify areas for improvement. As mentioned before, to evaluate our system, we will compare our results to the ones obtained by the QTM. This section details the evaluation process, particularly focusing on the metrics chosen.

6.2.1 Evaluation Metrics

In order to evaluate such a system, we must first consider the task we are performing and what evaluation metrics better suit it. In this case, we are predicting a score for the quality of a given rehabilitation exercise based on measuring distance. In more practical terms, we are predicting a continuous value for each entry in our dataset, which, despite not being a regression problem, can be evaluated with typical regression metrics. With that in mind, we selected Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Pearson Correlation Coefficient (CC) to evaluate our system.

6.2.1.1 Mean Absolute Error

MAE measures the average absolute difference between the predicted scores and the actual scores. It provides an interpretable metric for understanding the overall accuracy of our system's predictions.

Equation 6.4 shows how to calculate MAE, with N being the number of predictions, y_i being the actual values, and $\hat{y}_i$ being the predicted values.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{6.4}$$

6.2.1.2 Root Mean Squared Error

RMSE calculates the square root of the average squared difference between the predicted scores and the actual scores. It penalizes larger errors more significantly than MAE and provides insight into the magnitude of prediction errors.

Equation 6.5 is the RMSE equation, with N, y_i, and $\hat{y}_i$ having the same meaning as in the previous section.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{6.5}$$

6.2.1.3 Correlation Coefficient

CC assesses the linear relationship between the predicted scores and the actual scores. It quantifies the strength and direction of the linear association, providing a measure of how well the variation in one variable explains the variation in another. In our case, it tells us how well the variation in the predicted scores explains the variation in the actual scores.

Equation 6.6 presents the formula to compute CC, where N, y_i, and $\hat{y}_i$ keep their meaning and m and $\hat{m}$ are the mean of y and $\hat{y}$, respectively.

$$CC = \frac{\sum_{i=1}^{N}(y_i - m)(\hat{y}_i - \hat{m})}{\sqrt{\sum_{i=1}^{N}(y_i - m)^2 \cdot \sum_{i=1}^{N}(\hat{y}_i - \hat{m})^2}} \tag{6.6}$$

6.2.2 Outliers

Before we show the results obtained with the chosen metrics, mentioning outliers in our dataset is very important. As mentioned in Section 5.3, the dataset used in our work contains outliers that have been labeled as such by the researchers who collected the data. This means we have to handle them correctly, as they can have strong implications in our evaluation process.

One way to handle outliers is to remove them from the data directly [2], making it so that they would not affect our evaluation. However, our data is multidimensional, as we have a dimension for each joint group, Head, Trunk, and Shoulder, and the outliers appear only in one dimension. In the case of the *Diagonal* Exercise, the Head measurement of patient ID15 was considered an outlier. For the *Rotation* Exercise, the Head measurements of patients ID14 and ID15 were considered outliers. If we opt to remove the whole exercise from the dataset, we remove the outlier successfully, but we also remove valid data, more specifically, the measurements of the other joint groups of the mentioned patients.

Given this scenario, we could impute the value of the outliers. Imputation is a technique used to fill in missing data or handle outliers in a dataset by, for example, replacing them with the mean of the other values [11]. This would allow us to evaluate the complete dataset. Another alternative would be to remove the values just from the mentioned joint groups and evaluate the other joint groups for those patients. We ended up evaluating the joint groups separately, so it made sense to remove just those values instead of imputing them.

6.3 Results

After analyzing our data, normalizing it, and selecting the best metrics, we are ready to evaluate our system. This section presents the results of the evaluation, in which the system's performance is assessed through a comparison with the QTM. This comparison allows us to evaluate our system's efficacy against a very powerful tracking methodology. The results will be presented in two separate sections, one for each exercise. Each section has three subsections, one for each joint group. Each subsection includes scores normalized using Z-score normalization and the scores normalized using Min-max scaling, as mentioned in Section 6.1.

6.3.1 *Diagonal* Exercise

This section contains the results of the *Diagonal* exercise, divided into the three joint groups. For each joint group, we present the results obtained from the selected evaluation metrics: MAE, RMSE, and CC.

6.3.1.1 Head

The values obtained by the metrics for both the Head joint group of the *Diagonal* exercise are shown in Table 6.2 and Figure 6.3 shows the error distribution for the Min-max predictions and the Z-score predictions.

Normalization	**MAE**	**RMSE**	**CC**
Min-max	13.29	15.52	0.36
Z-score	27.88	34.36	

Table 6.2: Evaluation metrics for the Head joint group for the *Diagonal* exercise

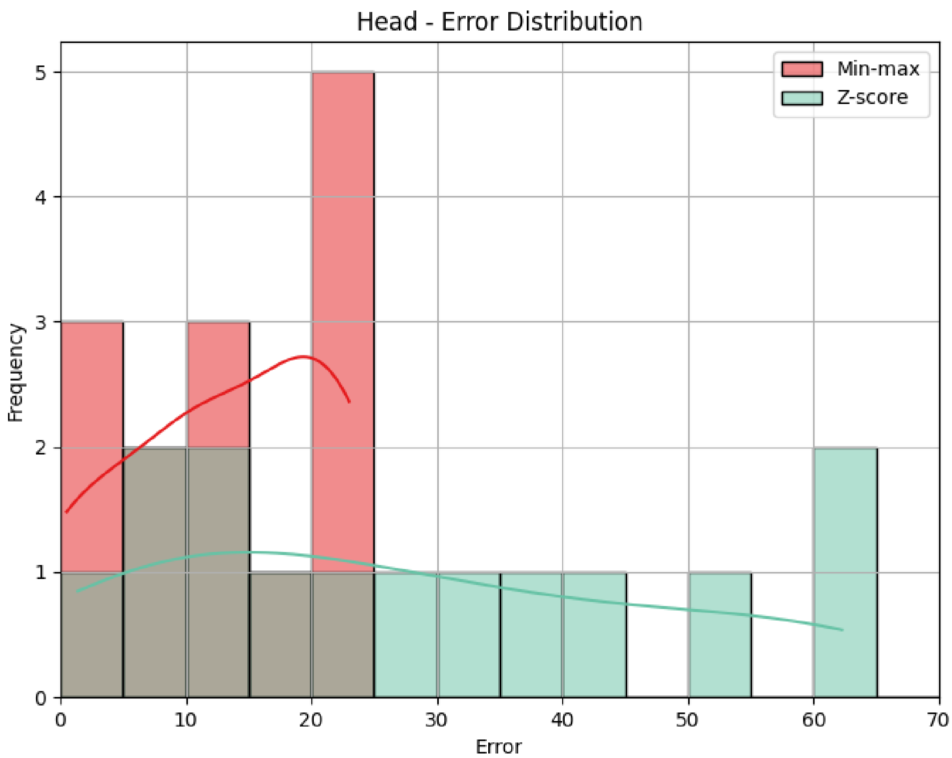

Figure 6.3: Error distribution for the Head joint group for the *Diagonal* exercise

Using Min-max scaling, the Head joint group for the *Diagonal* exercise obtained an MAE of 13.29, an RMSE of 15.52, and a CC of 0.36.

The value of the MAE tells us how big were the errors in our prediction, on average, in the scale of our results, that is, 0 to 100. This means that the Head joint group's predicted score was, on average, 13.29 points away from the score obtained by the QTM.

The RMSE value calculates the difference between the squares of the errors, penalizing larger errors. In our case, the values are not too different, meaning that most errors are small and that there was little influence by large errors.

The CC of 0.36 indicates a positive but weak correlation, as 0.36 is positive but small compared to the ideal value of 1. The interpretation of this value can be that the variation of our predictions is in the same direction as the actual values, meaning when our predictions increase, the actual values tend to increase as well, and vice-versa, but that relationship is not strong, meaning not much of the variation in the actual values can be explained by the variation in our predictions.

When it comes to the Z-score results' evaluation, if we compare the MAE and the RMSE to those of the Min-max results, we can immediately detect that Z-score caused more errors, as the MAE is more than double, and larger errors, with an even larger difference of RMSE. This is expected, given what was explained in Section 6.1: Z-score normalization is not necessarily suitable for our data. The difference in errors can be visually understood with Figure 6.3. The CC is the same.

6.3.1.2 Trunk

The values obtained by the metrics for the Trunk joint group of the *Diagonal* exercise are shown in Table 6.3. In Figure 6.4, we can see the error distribution for both sets of predictions.

Normalization	**MAE**	**RMSE**	**CC**
Min-max	17.10	19.12	0.47
Z-score	31.11	37.67	

Table 6.3: Evaluation metrics for the Trunk joint group for the *Diagonal* exercise

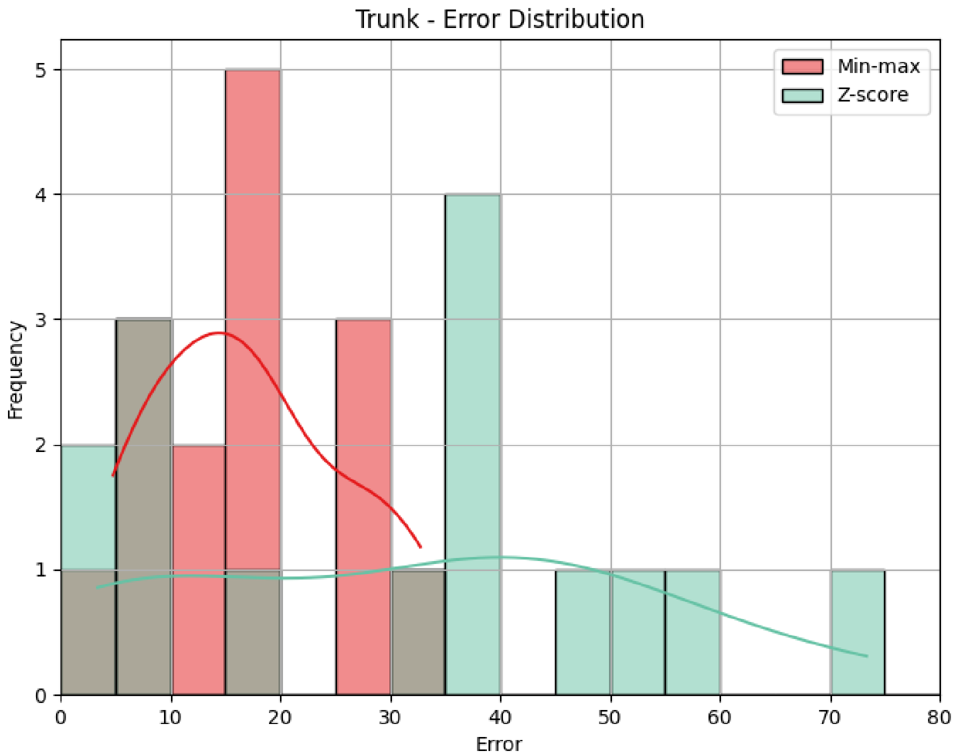

Figure 6.4: Error distribution for the Trunk joint group for the *Diagonal* exercise

With Min-max scaling, the Trunk joint group obtained an MAE of 17.10, an RMSE of 19.12, and a CC of 0.47.

The MAE of 17.10 reveals that, on average, the predicted scores for the Trunk joint group deviated by 17.10 points from the QTM scores. As for the Head, while there are some larger errors, the majority are smaller, as indicated by similar values of MAE and RMSE.

The CC of 0.47 indicates a positive correlation, and it is relatively stronger than observed in the Head joint group. The variation in our predictions aligns with the variation in actual values, suggesting a reasonably strong directional relationship.

In contrast, the Z-score results for the Trunk joint group show a higher MAE of 31.11 and a larger RMSE of 36.67 compared to Min-max scaling. This notable increase in errors in magnitude and quantity reveals the difference between Min-max and Z-score and can be better understood when we look at Figure 6.4. Again, the CC value is the same.

6.3.1.3 Shoulder

The evaluation metrics for the Shoulder joint group during the *Diagonal* exercise are illustrated in Table 6.4 and the error distribution for both sets of predictions is represented in Figure 6.4.

Normalization	**MAE**	**RMSE**	**CC**
Min-max	14.44	15.59	-0.37
Z-score	57.59	62.97	

Table 6.4: Evaluation metrics for the Shoulder joint group for the *Diagonal* exercise

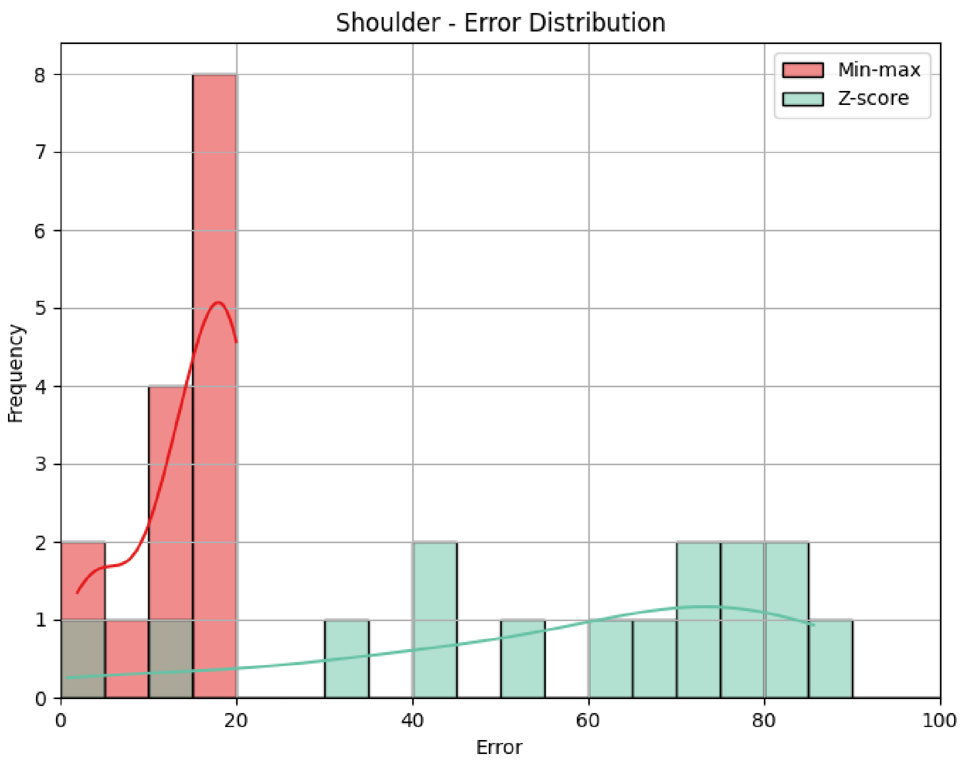

Figure 6.5: Error distribution for the Shoulder joint group for the *Diagonal* exercise

Under Min-max scaling, the Shoulder joint group obtained an MAE of 14.44, an RMSE of 15.59, and a CC of -0.37.

The MAE is similar to the one obtained by the Head and represents an average deviation of 14.44 points from the QTM scores. The values of MAE and RMSE suggest a reasonable level of accuracy, with few large errors, given how close the two values are.

The CC of -0.37 indicates a negative correlation, which means that when the values in our predictions increase, the values in the actual values decrease, which is not ideal. However, this correlation is weak, which means not much of the variation in the actual values can be explained by the variation in our predictions.

Following the trend of the other two joint groups, the Z-score results for the Shoulder joint group show a substantially higher MAE of 57.59 and a larger RMSE of 62.97. In this case, the difference is even greater, meaning a considerable increase in errors, as can be seen in Figure 6.5.

6.3.2 *Rotation* Exercise

6.3.2.1 Head

The values obtained by the metrics for the Head joint group of the *Rotation* exercise are shown in Table 6.5. Figure 6.6 shows the error distribution for the Head joint group during the *Rotation* exercise.

Normalization	**MAE**	**RMSE**	**CC**
Min-max	20.41	24.60	0.22
Z-score	18.86	27.11	

Table 6.5: Evaluation metrics for the Head joint group for the *Rotation* exercise

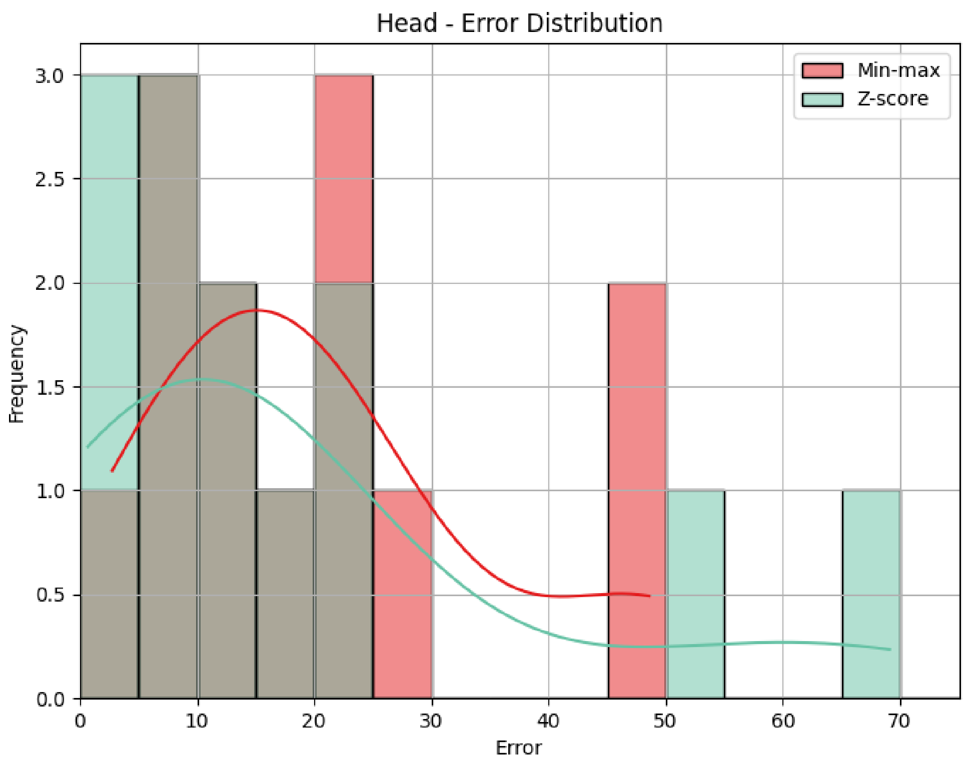

Figure 6.6: Error distribution for the Head joint group for the *Rotation* exercise

When employing Min-max scaling, the Head joint group obtained an MAE of 20.41, an RMSE of 24.60, and a CC of 0.22.

The MAE of 20.41 implies that, on average, the predicted scores for the Head joint group deviated by 20.41 points from the QTM scores. The RMSE, measuring the square root of the average squared errors, is slightly larger, indicating the presence of some large errors.

The CC of 0.22 indicates a positive but weak correlation. This suggests a directional relationship between our predictions and the actual values, but the correlation strength is low.

Utilizing Z-score normalization, the Head joint group achieved an MAE of 18.86 and an RMSE of 27.11. While the MAE shows a slight improvement in prediction accuracy compared to Min-max scaling, the RMSE reflects an increase in the magnitude of errors. In Figure 6.6, we can see that Z-score shows, in general, smaller errors than Min-max, explaining the lower MAE, but, as we can see, it also shows one very large error, penalized by RMSE. As mentioned in the previous sections, the CC is the same for Z-score and Min-max results.

6.3.2.2 Trunk

Table 6.6 displays the evaluation metrics for the Trunk joint group during the 'Rotation' exercise and in Figure 6.7, we can see the error distribution for the Trunk joint group during the *Rotation* exercise.

Normalization	**MAE**	**RMSE**	**CC**
Min-max	32.30	33.69	0.51
Z-score	27.30	32.25	

Table 6.6: Evaluation metrics for the Trunk joint group for the *Rotation* exercise

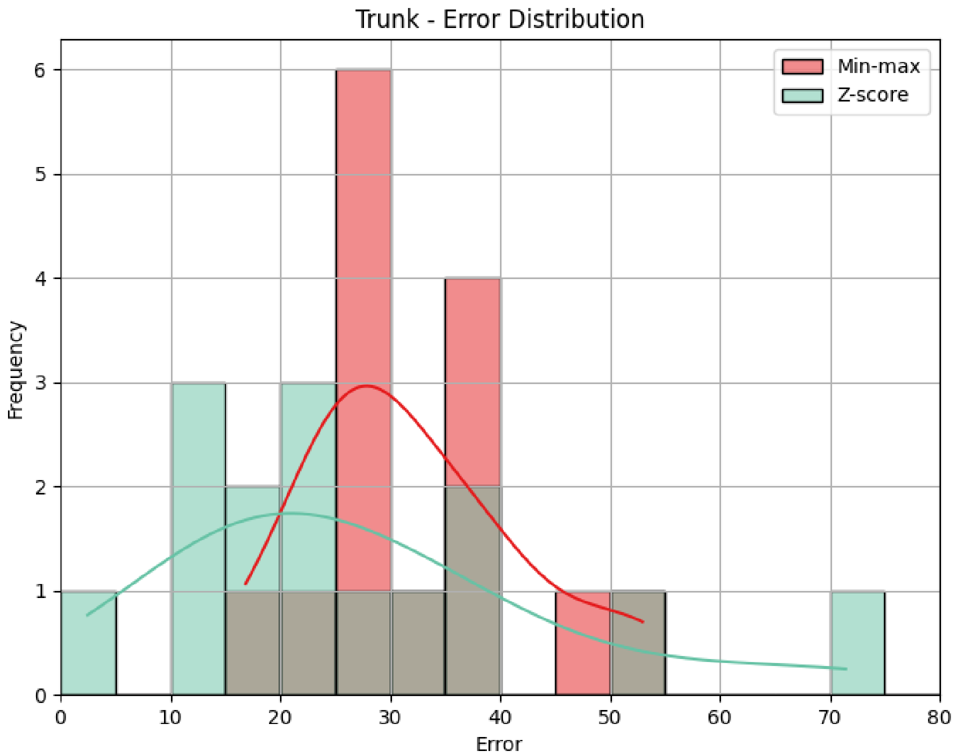

Figure 6.7: Error distribution for the Trunk joint group for the *Rotation* exercise

With Min-max scaling, the Trunk joint group achieved an MAE of 32.30, an RMSE of 33.69, and a CC of 0.51.

The MAE of 32.30 is the largest of any joint group with Min-max, indicating larger errors on average. A similar value for RMSE reveals that the largest error must not be much larger than the other errors. However, because the MAE is already large, the RMSE could still indicate the presence of large errors.

The CC of 0.51 reveals a positive correlation with a moderate strength. This suggests a reasonably strong directional relationship between our predictions and the actual values for the Trunk joint group.

Switching to Z-score normalization, the Trunk joint group achieved an MAE of 27.30 and an RMSE of 32.25. Notably, Z-score normalization resulted in a decrease of the MAE, indicating an improvement in prediction accuracy, while the value of the RMSE being close to the one obtained with the Min-max results implies a comparable distribution of errors. Figure 6.7 shows that Z-score presents many more errors close to 0, but there are still some very large errors, while Min-max's errors are generally larger. Once more, the CC is the same.

6.3.2.3 Shoulder

The evaluation metrics for the Shoulder joint group during the 'Rotation' exercise are illustrated in Table 6.7. Figure 6.8 displays the error distribution for the Trunk joint group during the *Rotation* exercise.

Normalization	MAE	RMSE	CC
Min-max	21.45	23.19	-0.37
Z-score	30.30	36.83	

Table 6.7: Evaluation metrics for the Shoulder joint group for the *Rotation* exercise

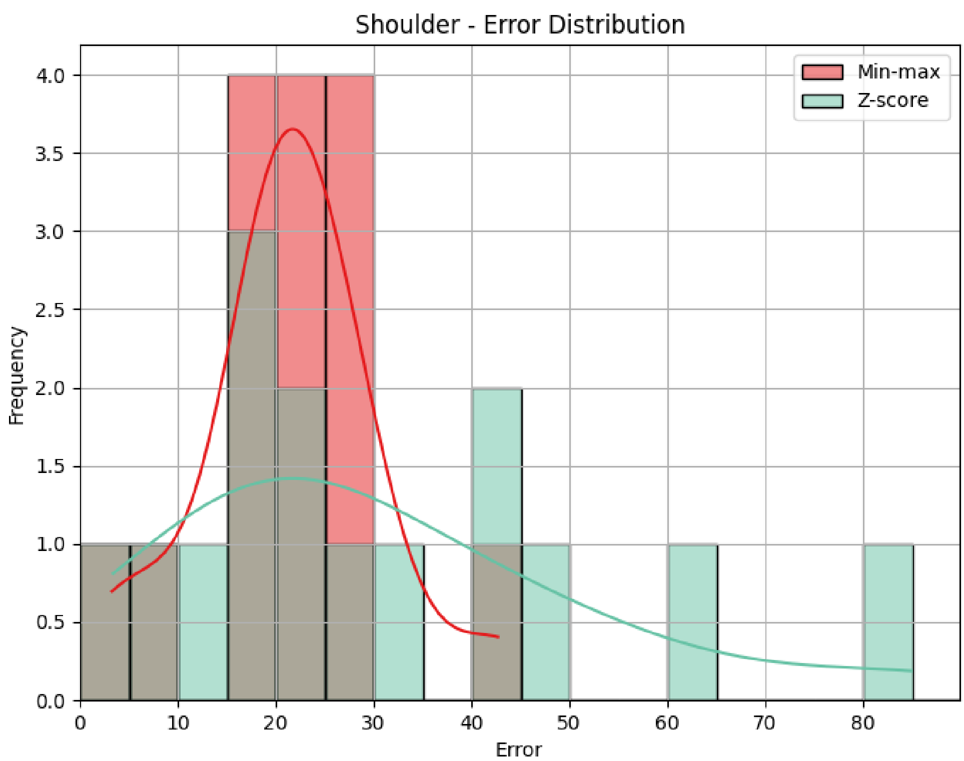

Figure 6.8: Error distribution for the Shoulder joint group for the *Rotation* exercise

When employing Min-max scaling, the Shoulder joint group obtained an MAE of 21.45, an RMSE of 23.19, and a CC of -0.37.

Once again, the value of the MAE indicates that, on average, the predicted scores for the Shoulder joint group deviated by 21.45 points from the QTM scores. The RMSE suggests reasonably large errors, as it is larger than MAE.

As for the Shoulder group of the *Diagonal* exercise, the CC of -0.37 indicates a negative correlation, revealing an inverse relationship between our predictions and the actual values for the Shoulder joint group. However, this correlation is weak, with a relatively low magnitude.

Upon employing Z-score normalization, the Shoulder joint group achieved an MAE of 30.30 and an RMSE of 36.83 Z-score normalization resulted in an increase in both the MAE, indicating a decrease in prediction accuracy, and the RMSE, revealing the presence of larger errors. Figure 6.8 clarifies the difference between MAE and RMSE values. Min-max shows not only smaller values on average but also larger values that are smaller than the ones obtained by Z-score. Once again, the value of the CC is not affected by normalization.

6.4 Results Discussion

The presented results shed light on the performance of our system in predicting the quality scores of rehabilitation exercises compared to the QTM. In this section, we discuss key findings and potential implications of the obtained results.

Before we take a longer look at the results, however, it is relevant to state that the choice of normalization method significantly impacts our results. Min-max scaling consistently outperforms Z-score normalization, showcasing its suitability for our dataset, as we had clear minimum and maximum values. While Z-score normalization is widely adopted, our findings emphasize the importance of exploring alternative normalization techniques, as suggested by Lima and Souza [36].

It is also important to mention outlier handling and how crucial it is. Our decision was to remove values from specific joint groups rather than entire exercises, as this approach preserves valuable data while addressing the influence of outliers on specific joint measurements.

6.4.1 Overall Performance Trends

Across both the *Diagonal* and *Rotation* exercises, the system demonstrates varying levels of accuracy and correlation for different joint groups. That is noticeable when we look at the ranges of values obtained for each calculated error metric. When it comes to the Min-max scores, MAE ranged from 13.29 to 32.30, and RMSE ranged from 15.52 to 33.69. In both metrics, a difference of more than double is present. Z-score MAE ranged from 18.86 to 57.7, and RMSE ranged from 27.11 to 62.97. This tells us that performances were not consistent over different joint groups and exercises.

In terms of correlation, we obtained reasonable results overall, with most of the correlations being positive, including the Trunk group, which obtained CCs around the 0.5 mark. However, the Shoulder group evaluation resulted in negative coefficients, which indicates an opposite-direction relationship between the predicted and the actual values.

This means that, overall, there was a mixture of results, which means we must compare results between exercises and joint groups and discuss whether there are any indications of one exercise outperforming the other or one joint group outperforming the others.

6.4.2 Exercise-Specific Observations

This subsection includes observations for each exercise and joint group, allowing for comparison between them and a better understanding of our results.

6.4.2.1 *Diagonal* Exercise

Considering the Min-max results for the *Diagonal* exercise, the Head group presents modest errors and exhibits a positive correlation, which, despite being weak, suggests a consistent directional relationship. The Trunk group's errors are slightly larger and include a positive correlation, highlighting reasonable prediction accuracy. The Shoulder groups' evaluation metrics' results are between those of the Head and the Trunk groups. However, a negative correlation suggests that predictions may not be related to actual values. Overall, the Min-max scores for this exercise obtained reasonable results, with an average MAE of 14.94 and an average RMSE of 16.74. Although the Head had fewer errors, the values are too close to state a clear difference between any of the joint groups.

With the Z-score scores, the errors were larger. While the Head and Trunk groups obtained MAE and RMSE values close to 30, which would already be worse than the Min-max scores, the Shoulder group obtained an MAE and an RMSE close to 60, revealing that the average prediction was off by almost 60 points.

6.4.2.2 *Rotation* Exercise

Regarding the Min-max results for the *Rotation* exercise, the Head and Shoulder groups obtained similar error metrics, and the Trunk group obtained a higher degree of error. In terms of correlation, the trend from the *Diagonal* exercise continues, with the Head obtaining a weak positive correlation, the Trunk a reasonably high positive correlation, and the Shoulder a weak negative correlation. Compared with the *Diagonal* exercise, the Min-max scores for this exercise obtained worse results, with an average MAE of 24.72 and an average RMSE of 27.16, practically 10 points higher than the respective metrics for the *Diagonal* exercise.

Interestingly, the Z-score scores obtained very similar, if not better, performances to those of the Min-max scores in both the Head and the Trunk groups, obtaining slightly lower MAEs and similar RMSEs.

6.4.3 Discussion Summary

In summary, the evaluation of the rehabilitation exercise prediction system offers key insights into its performance across the *Diagonal* and *Rotation* exercises, focusing on joint groups—Head, Trunk, and Shoulder.

The results highlight the impact of data normalization, emphasizing the superior performance of Min-max scaling over Z-score normalization.

The chosen evaluation metrics - MAE, RMSE, and CC - help us understand the accuracy of the performance of our system and how related our predictions may be with the actual scores, unveiling performance trends across joint groups and exercises, such as the overall better performance of our system during the *Diagonal* exercise and the higher correlation results yielded from the Trunk group.

6.5 Statistical Analysis

After computing traditional evaluation metrics, we wanted to better understand our results through hypothesis testing. According to Whitley and Ball [59], hypothesis tests are divided into two groups: parametric tests and non-parametric tests. The main difference between the two groups is that parametric tests expect assumptions to be made about the data, most commonly requiring the data to follow a normal distribution. Alternatively, non-parametric tests make no assumptions about the data.

Given the nature of our data, characterized by a relatively small population and the inability to assume a specific distribution, we opted for non-parametric tests. More specifically, we opted for the Wilcoxon Signed-Rank Test [61], which is particularly suitable for paired samples, in our case the predicted and actual scores, providing a robust analysis of whether there is a significant difference between them.

6.5.1 Wilcoxon Signed-Rank Test

The Wilcoxon Signed-Rank Test [61] is a non-parametric statistical test used to assess whether there is a significant difference between paired samples. The basic idea is to rank the absolute differences between the pairs of observations, which in our case are the errors, ignoring the signs, and sum the ranks of the positive or negative differences. The null hypothesis assumes that the distribution of these differences is symmetric around zero. The test statistic is then calculated based on these ranks [59].

The steps for performing this test are as follows:

1. For each pair of observations, calculate the absolute difference between the predicted and the actual score
2. Rank these absolute differences from smallest to largest, ignoring the signs
3. Taking the signs back into account, sum all the positive ranks (R_-) and all the negative ranks (R_+)
4. Calculate the smallest value between R_- and R_+: $R = \min(R_-, R_+)$

The final value, R, is the test statistic and can be compared with tabled critical values to obtain the p-value. That p-value is used to assess the significance of the difference found between the two sets of scores. If it is under a given significance value, commonly 0.05, we can reject the null hypothesis and state that there is a significant statistical difference between the two sets of scores. If it is over that value, we cannot confirm that that statistical difference is significant.

Thus, we performed the Wilcoxon Signed-Rank Test for the differences between the Min-max scores and the actual scores and the differences between the Z-score scores and the actual scores, divided into the three joint groups. We performed the test once for each exercise.

Table 6.8 presents the results for the Wilcoxon Signed-Rank Test for the *Rotation* exercise.

Exercise	**Joint Group**	**P-Value Min-Max**	**P-Value Z-score**
Diagonal	Head	0.0574	0.00464
	Trunk	0.00061	0.00201
	Shoulder	0	0.000122
Rotation	Head	0.0574	0.3396
	Trunk	0	0
	Shoulder	0.0353	0.1070

Table 6.8: Wilcoxon Signed-Rank Test Results for Min-Max and Z-score Scores for the *Rotation* and *Diagonal* Exercises

One problem with this test was noticed immediately, as the P-values of the Trunk joint group of the *Rotation* exercise for both the Min-max and the Z-score scores are 0, as well as the P-value of the Shoulder group of the *Diagonal* exercise for the Min-max scores. This is due to the fact that, for those particular scores, all of the predictions were lower than the actual score. This can be confirmed by interpreting Figures 6.9 and 6.10, which display a scatter plot of the Min-max and Z-score predictions against the actual scores.

As we can see, all the predictions in Figure 6.9 are under the red dashed line that indicates a correct prediction, meaning that all the differences are negative. The same thing happens for the Min-max predictions in Figure 6.10.

Going back to how the Wilcoxon Signed-Rank Test works, it is easy to understand that if the differences are all positive or all negative, one of R_- or R_+ will be 0, and thus R will always be 0. In these cases, the Wilcoxon Signed-Rank Test does not provide meaningful information.

However, the other values are of interest, and we can interpret them following what was mentioned earlier. For the *Diagonal* exercise, the Min-max Head value is more than 0.05, indicating that, for that pair of sets of scores, we cannot reject the null hypothesis. That is, we cannot confirm that there is a significant statistical difference between the sets of scores that generated those P-values. The rest of the sets of scores had P-values less than 0.05, confirming that the difference between the paired sets is not by chance and holds statistical significance.

Regarding the *Rotation* exercise, the P-values more than 0.05 include the Min-max Head, Z-score Head, and Z-score Shoulder values. Again, we cannot state that the differences present are statistically significant for these pairs of sets of scores. For the Shoulder Min-max value,

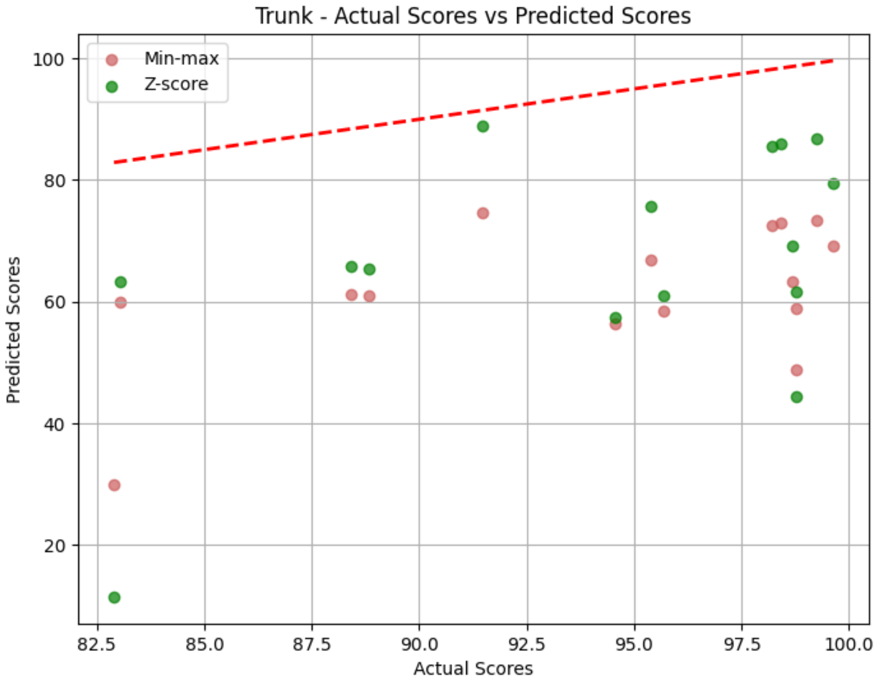

Figure 6.9: Scatter plot of predictions for the Trunk joint group for the *Rotation* exercise

which is less than 0.05, we reject the null hypothesis and confirm the statistical significance of the differences between the sets of scores.

6.6 Summary

This chapter focuses on evaluating our CV-based motion tracking system for rehabilitation. We began by discussing the importance of data normalization, specifically for comparing results obtained from a distance-based system with a ground truth dataset labeled on a 0 to 100 scale. Three normalization methods were presented: Z-score normalization, Min-max scaling, and Maximum absolute scaling. We mentioned that we opted to use Min-max scaling, based on the characteristics of the dataset, but also Z-score normalization, as it is very commonly used.

We then explored the system evaluation process, emphasizing the significance of choosing appropriate metrics. MAE, RMSE, and CC were selected to assess the system's prediction accuracy. We also addressed the outliers in the dataset and how we handled them by removing specific joint group measurements rather than entire exercises.

Then, we presented the results for the *Diagonal* and *Rotation* exercises, focusing on the Head, Trunk, and Shoulder joint groups. We discussed the results, highlighting the impact of data normalization and how Min-max scaling outperformed Z-score normalization for the most part. We touched on overall performance trends that showed varying levels of accuracy across the different exercises and joint groups. Thus, we included exercise-specific observations, revealing that the

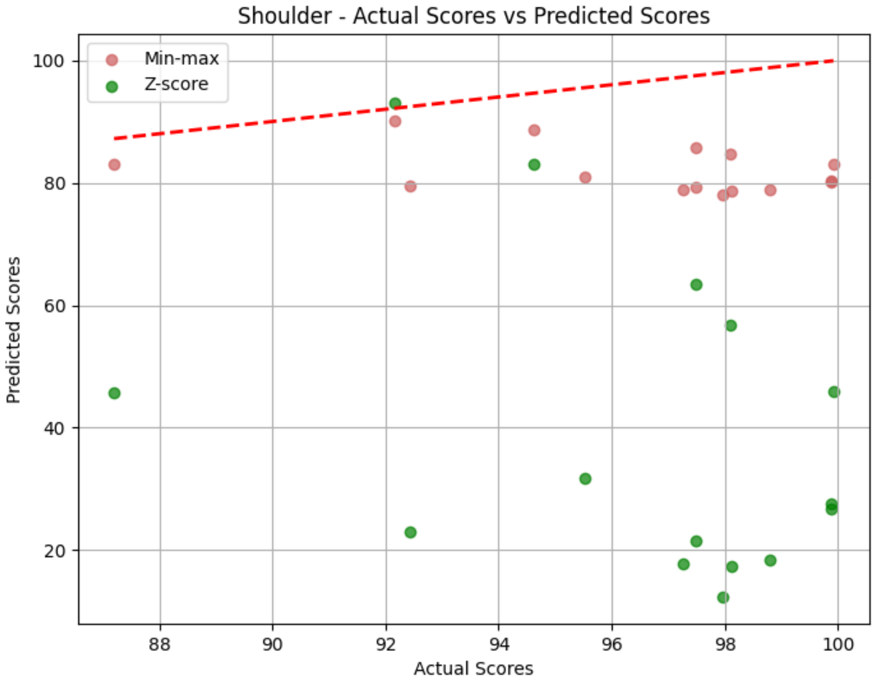

Figure 6.10: Scatter plot of predictions for the Shoulder joint group for the *Diagonal* exercise

system performed better during the *Diagonal* exercise and that the Trunk joint group obtained higher overall correlation results.

After evaluation, we employed hypothesis testing to gain insights into the evaluation metrics obtained earlier. We introduced parametric and non-parametric tests and explained the difference between the two. In our case, there was a preference for non-parametric methods due to the dataset's characteristics, such as a small population and the absence of assumptions about data distribution. The Wilcoxon Signed-Rank Test was specifically chosen for its applicability to paired samples, comparing predicted and actual scores. The test was explained, and the steps for performing it were outlined. However, a limitation with the test arose, as when all predictions are consistently lower than actual scores, the test is rendered ineffective. Despite this limitation, the chapter provides valuable insights into the statistical significance of differences between predicted and actual scores for various exercise and joint group combinations.

Chapter 7

Conclusion

This chapter marks the conclusion of this dissertation, reviewing the work that has been done and the challenges faced in Section 7.1. A summary of the main takeaways of this project is presented in Section 7.2. Section 7.3 touches on the main opportunities for improvement of this project and future work possibilities. Lastly, Section 7.4 includes a reflection on the dissertation as a whole, discussing the overall balance of this journey.

7.1 Recap

In this thesis, we have explored the fields of computer vision systems and rehabilitation and how the former can be applied to improve the latter and make it more accessible.

Our main goal, as mentioned in Section 1.3, was to develop a Computer-Vision-based system that allows patients to perform rehabilitation exercises independently, using accessible technology, while receiving necessary and adequate feedback. In order to do that, we had to research the two main fields involved: Rehabilitation and Computer Vision.

We began by understanding the rehabilitation context in Chapter 1, getting a better grasp of the current state of rehabilitation and how it has evolved historically to get to the present.

Then, in Chapter 2, we established fundamental concepts of Computer Vision, including relevant CV tasks, and Deep Learning, mentioning relevant models, such as different types of Neural Networks.

After that, in Chapter 3, we examined the state of the art in Computer Vision, focusing on its relevance to accessible rehabilitation. We conducted an in-depth analysis of existing computer vision systems for intelligent rehabilitation. This meant researching and reviewing the literature to understand what current solutions were offered and how the researchers viewed and solved this problem. We also tried to understand what the most accessible options were in terms of technology to be used.

After that, in Chapter 4, we identified gaps and opportunities for improvement and stated the problem at hand.

We then explained the implementation of our solution in Chapter 5. This was the moment of our journey when theory was put into practice, as we took what we learned from our research and applied it to our solution. Building on what the literature review had taught us about the typical architecture of systems similar to the one we were about to design, we opted to follow the same structure, which included *Data Acquisition* followed by *Feature Engineering* and then *Comparison and Assessment*. For each of those three components, we made choices based on accessibility, as was justified in Section 5.2. While the choices made for *Data Acquisition* and *Feature Engineering* were clear, for the *Comparison and Assessment* component, the choice was not easy, and it would have been of value to explore other alternatives. We will discuss this in more detail in Section 7.3.

We then analyzed the dataset made available to us and realized that it included 90 videos, comprising 15 patients performing two exercises three times each. However, the reference scores in the dataset were not attributed to each example in the dataset but to each patient, meaning we had a total of 30 scores: two for each of the 15 patients. This meant that, in general, the dataset was not very big.

With this in mind, we devised a plan to implement our system. Starting by simply comparing two videos, we immediately faced some challenges. For example, we started by making it so that the time sequence of coordinates of each joint group was represented by an array. This meant that each array would be composed of smaller arrays, one for each frame, and each one of those would be composed of more arrays, one for each set of X, Y, and Z coordinates of each joint relevant to that joint group. This meant the arrays had three dimensions, which the DTW libraries used had trouble with. Thus, we separated the arrays, retrieving one array for each relevant joint and calculating DTW over those arrays, summing each joint relevant for each joint group to obtain the distance of that group.

Another challenge faced was the time it took to process each video. Due to the joint tracking operations being performed, the video processing became very slow, which was very challenging, particularly when we went from comparing two videos to comparing the entire dataset. It would take a long time to go over the dataset to obtain the desired distance measures, meaning that any change affecting the results would imply a longer wait.

Finally, we evaluated our system and showed and analyzed the results obtained in Chapter 6. This process also involved some research to identify the adequate data normalization methods, the most suitable evaluation metrics, and how to best analyze the data statistically. We then performed that evaluation and discussed the results.

The implemented system builds on the current body of knowledge and presents an alternative solution, particularly focused on tackling the biggest problem faced by the field of rehabilitation: accessibility.

7.2 Summary of Main Takeaways

In the course of this research, we made pivotal observations and conclusions, from which we want to highlight the most important ones.

Firstly, we reviewed the current state of the art in Computer Vision systems for assessing physical rehabilitation exercises. This work resulted in the presentation and publication of a scientific article [40].

Then, we successfully built a CV-based motion tracking system while prioritizing accessibility. That means we successfully demonstrated that there is accessible technology currently available to fill this gap. This could indicate a shift toward developing rehabilitation systems based on accessible technology. One of the aspects of our work that greatly supports that shift is the fact that the smartphone can be used to successfully capture rehabilitation-relevant data. As has been mentioned before, we agree with Lam et al. [31] that smartphones will play a pivotal role in the future of rehabilitation.

Another important takeaway is that DTW can be used to compare two videos. Despite the trend being the integration of Deep Learning, we are in agreement with the current state of research that DTW can be used as a benchmark.

Finally, we also confirm that Z-score normalization does not suit non-normally distributed data. Our findings underscore the significance of choosing appropriate normalization methods, with Min-max scaling emerging as a preferred alternative in scenarios where there is a clear maximum and minimum defined.

These takeaways encapsulate the key achievements and insights derived from our research, serving as a brief summary of what can be concluded from this dissertation, research-wise.

7.3 Opportunities for Improvement and Future Work

The final result of this dissertation can still be improved, which is shown by the results being worse than desired.

One of the main opportunities for improvement, as mentioned in Section 7.1, is the lack of exploration of alternatives for the *Comparison and Assessment* component of our system. More specifically, as understood from Chapter 3, while DTW shows interesting results, the trend is to explore Deep Learning. Moving forward, this should be the main focus of future work. Its integration into the project was hindered by time constraints and the fact that the dataset available was not particularly big.

The size of the dataset could also be improved, as more data could be either retrieved or artificially generated. A more expansive and balanced dataset would not only bolster the generalizability of the system but also facilitate the effective training of complex Deep Learning models.

Another aspect to consider is the ethical outlines of this work. In a possible integration in an application made available to patients, a comprehensive exploration of ethical guidelines should be performed, prioritizing patient privacy. This could include collaboration with professionals to align technological advancement with ethical considerations. Efforts should be made to keep the patients safe.

Acknowledging limitations serves as a stepping stone for future advancements. The outlined areas for future work preview how the system could be improved, aligning it with research trends and upholding ethical standards.

7.4 Final Remarks

As this journey comes to an end, we reflect on the work done and the contributions made.

Overall, we believe that this project has fulfilled its goal of providing an accessible tool for rehabilitation based on Computer Vision. By designing this CV system for rehabilitation, we have successfully laid the groundwork for the future of accessible rehabilitation.

Beyond the technical work described throughout this dissertation, there was a personal component that was pivotal to this journey. From the initial project proposal to the literature review and understanding of the problem to then developing and evaluating our system, this process proved challenging in several ways. This has a very positive effect, leading to the development of problem-solving skills and persevering through difficulties.

In conclusion, this project signifies more than an academic task. It is proof that when we combine sophisticated technology with determination, we can make a meaningful difference. We conclude this journey with a sense of fulfillment in having improved the realm of rehabilitation.

References

[1] Anzar Abbas, Vijay Yadav, Emma Smith, Elizabeth Ramjas, Sarah B Rutter, Caridad Benavidez, Vidya Koesmahargyo, Li Zhang, Lei Guan, Paul Rosenfield, Mercedes Perez-Rodriguez, and Isaac R Galatzer-Levy. Computer vision-based assessment of motor functioning in schizophrenia: Use of smartphones for remote measurement of schizophrenia symptomatology. *Digital Biomarkers*, 5(1):29–36, 2021.

[2] Herman Aguinis, Ryan K. Gottfredson, and Harry Joo. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2):270–301, 2013.

[3] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016.

[4] Sandra Bassett and Harry Prapavessis. Home-based physical therapy intervention with adherence-enhancing strategies versus clinic-based management for patients with ankle sprains. *Physical therapy*, 87:1132–43, 2007.

[5] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv pre-print server*, 2020.

[6] Chris M. Bishop. Neural networks and their applications. *Review of Scientific Instruments*, 65(6):1803–1832, 1994.

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *arXiv pre-print server*, 2019.

[9] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriù, L. Romeo, and F. Verdini. The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7):1436–1448, 2019.

[10] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Ville Kyrki, Andrea Monteriù, Luca Romeo, and Federica Verdini. A hidden semi-markov model based approach for rehabilitation exercise assessment. *Journal of Biomedical Informatics*, 78:1–11, 2018.

[11] Raymond L. Chambers and Ruilin Ren. Outlier robust imputation of survey data. *The Proceedings of the American Statistical Association*, pages 3336–3344, 2004.

[12] Yen-Lin Chen, Chin-Hsuan Liu, Chao-Wei Yu, Posen Lee, and Yao-Wen Kuo. An upper extremity rehabilitation system using efficient vision-based action identification techniques. *Applied Sciences*, 8(7):1161, 2018.

[13] Lucio Ciabattoni, Francesco Ferracuti, Sabrina Iarlori, Sauro Longhi, and Luca Romeo. A novel computer vision based e-rehabilitation system: From gaming to therapy support. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*, pages 43–44, 2016.

[14] Steffi L. Colyer, Murray Evans, Darren P. Cosker, and Aki I. T. Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Medicine - Open*, 4(1), 2018.

[15] Bappaditya Debnath, Mary O'Brien, Motonori Yamaguchi, and Ardhendu Behera. A review of computer vision-based approaches for physical rehabilitation and assessment. *Multimedia Systems*, 28(1):209–239, 2022.

[16] Javier Dorado, Xavier del Toro Garcia, Maria Santofimia, Alfonso Parreño-Torres, Ruben Cantarero, Ana Rubio Ruiz, and Juan Carlos López. A computer-vision-based system for at-home rheumatoid arthritis rehabilitation. *International Journal of Distributed Sensor Networks*, 15:155014771987564, 2019.

[17] Shradha Dubey and Manish Dixit. A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29(1):167–195, 2023.

[18] Exer. Exer health. https://www.exer.ai/health, 2023. Accessed: 2023-05-30.

[19] Soroush Falahati. *OpenNI Cookbook*. Packt Publishing, 2013.

[20] Elisa Ferrer-Mallol, Clare Matthews, Madeline Stoodley, Alessandra Gaeta, Elinor George, Emily Reuben, Alex Johnson, and Elin Haf Davies. Patient-led development of digital endpoints and the use of computer vision analysis in assessment of motor function in rare diseases. *Frontiers in Pharmacology*, 13, 2022.

[21] João A. Francisco and Paulo Sergio Rodrigues. Computer vision based on a modular neural network for automatic assessment of physical therapy rehabilitation activities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pages 1–1, 2022.

[22] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[23] Google. Mediapipe. https://developers.google.com/mediapipe, 2023. Accessed: 2023/12/21.

[24] Google. Pose landmark detection guide. https://developers.google.com/mediapipe/solutions/vision/pose_landmarker, 2023. Accessed: 2023/12/21.

[25] Thomas Hellsten, Jonny Karlsson, Muhammed Shamsuzzaman, and Göran Pulkkis. The potential of computer vision-based marker-less human motion analysis for rehabilitation. *Rehabilitation Process and Outcome*, 10:117957272110223, 2021.

[26] Xianwei Huang, Fazel Naghdy, Golshah Naghdy, Haiping Du, and Catherine Todd. Robot-assisted post-stroke motion rehabilitation in upper extremities: a survey. *International Journal on Disability and Human Development*, 16(3):233–247, 2017.

[27] Cong-Cong Huo, Ya Zheng, Wei-Wei Lu, Teng-Yu Zhang, Dai-Fa Wang, Dong-Sheng Xu, and Zeng-Yong Li. Prospects for intelligent rehabilitation techniques to treat motor dysfunction. *Neural Regeneration Research*, 16(2), 2021.

[28] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. *arXiv pre-print server*, 2016.

[29] Ravi Komatireddy. Quality and quantity of rehabilitation exercises delivered by a 3-d motion controlled camera: A pilot study. *International Journal of Physical Medicine &; Rehabilitation*, 02(04), 2014.

[30] Neha Kulkarni. Effect of dynamic time warping using different distance measures on time series classification. *International Journal of Computer Applications*, 179:34–39, 2017.

[31] Winnie W. T. Lam, Yuk Ming Tang, and Kenneth N. K. Fong. A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 20(1), 2023.

[32] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[33] Nudpakun Leechaikul and Siam Charoenseang. *Computer Vision Based Rehabilitation Assistant System*, pages 408–414. Springer International Publishing, 2021.

[34] Michael H. Li, Tiago A. Mestre, Susan H. Fox, and Babak Taati. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *Journal of NeuroEngineering and Rehabilitation*, 15(1), 2018.

[35] Y. Liao, A. Vakanski, and M. Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020.

[36] Felipe Tomazelli Lima and Vinicius M. A. Souza. A large comparison of normalization methods on time series. *Big Data Research*, 34:100407, 2023.

[37] Maria Lopes, Ana S. C. Melo, Bruno Cunha, and Andreia S. P. Sousa. Smartphone-based video analysis for guiding shoulder therapeutic exercises: Concurrent validity for movement quality control. *Applied Sciences*, 13(22):12282, 2023.

[38] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.

[39] Steven Machlin, Julia Chevan, William Yu, and Marc Zodet. Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults. *Physical therapy*, 91:1018–29, 2011.

[40] José Maçães, Bruno Cunha, and Ivone Amorim. Computer vision for accessible intelligent rehabilitation: An overview. In *23rd International Conference on Hybrid Intelligent Systems (HIS'23)*. Springer, 2023.

[41] Microsoft. Kinect. https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows, 2022. Accessed: 2023-05-30.

[42] Alina Miron, Noureddin Sadawi, Waidah Ismail, Hafez Hussain, and Crina Grosan. Intellirehabds (irds)—a dataset of physical rehabilitation movements. *Data*, 6(5):46, 2021.

[43] Quentin Mourcou, Anthony Fleury, Bruno Diot, Céline Franco, and Nicolas Vuillerme. Mobile phone-based joint angle measurement for functional assessment and rehabilitation of proprioception. *BioMed Research International*, 2015:1–15, 2015.

[44] Hossein Mousavi Hondori and Maryam Khademi. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of Medical Engineering*, 2014:1–16, 2014.

[45] Bruno Pereira, Bruno Cunha, Paula Viana, Maria Lopes, Ana S. C. Melo, and Andreia S. P. Sousa. A machine learning app for monitoring physical therapy at home. *Sensors*, 24(1):158, 2023.

[46] PORTIC. Smarthealth. https://www.portic.ipp.pt/projects.html, 2023. Accessed: 2023/06/26.

[47] AB Qualisys. Qualisys track manager user manual. *Gothenburg: Qualisys AB*, 2006.

[48] Tammy Richmond, Christopher Peterson, Jana Cason, Mike Billings, Evelyn Abrahante Terrell, Alan Chong W. Lee, Michael Towey, Bambang Parmanto, Andi Saptano, Ellen R. Cohn, and David Brennan. American telemedicine association's principles for delivering telerehabilitation services. *International Journal of Telerehabilitation*, 9(2):63–68, 2017.

[49] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[50] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[51] Stan Salvador and Philip Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. In *KDD workshop on mining temporal and sequential data*, volume 6, pages 70–80. Seattle, Washington, 2004.

[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2019.

[53] Chuan-Jun Su, Chang-Yu Chiang, and Jing-Yan Huang. Kinect-enabled home-based rehabilitation system using dynamic time warping and fuzzy logic. *Applied Soft Computing*, 22:652–666, 2014.

[54] Richard Szeliski. Computer vision: Algorithms and applications. *Texts in Computer Science*, 2011.

[55] Lili Tao, Adeline Paiement, Dima Damen, Majid Mirmehdi, Sion Hannuna, Massimo Camplani, Tilo Burghardt, and Ian Craddock. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Computer Vision and Image Understanding*, 148:136–152, 2016.

[56] Aleksandar Vakanski, Hyung-pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. *Data*, 3(1):2, 2018.

[57] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018:1–13, 2018.

[58] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *arXiv pre-print server*, 2016.

[59] Elise Whitley and Jonathan Ball. Statistics review 6: Nonparametric methods. *Critical Care*, 6(1364-8535 (Print)):509, 2002.

[60] WHO. Rehabilitation. https://www.who.int/news-room/fact-sheets/detail/rehabilitation, 2023. Accessed: 2023-05-30.

[61] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[62] Renjie Wu and Eamonn Keogh. Fastdtw is approximate and generally slower than the algorithm it approximates. *IEEE Transactions on Knowledge and Data Engineering*, PP:1–1, 2020.

[63] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020.

[64] H. Yang, Y. Wang, and Y. Shi. Rehabilitation training evaluation and correction system based on blazepose. In *2022 IEEE 4th Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pages 27–30, 2022.

[65] Ying Xuan Zhi, Michelle Lukasik, Michael H. Li, Elham Dolatabadi, Rosalie H. Wang, and Babak Taati. Automatic detection of compensation during robotic stroke rehabilitation therapy. *IEEE Journal of Translational Engineering in Health and Medicine*, 6:1–7, 2018.