The Feasibility of AI Cameras as a Platform for Autonomous, Real-Time Rehabilitation with Performance Feedback

Charli Posner

MAY 5, 2023

FINAL YEAR PROJECT THESIS SUBMITTED IN SUPPORT OF THE DEGREE OF MASTER OF ENGINEERING IN COMPUTER SCIENCE & ELECTRONICS

Department of Electrical & Electronic Engineering
University of Bristol

1 Abstract

Because of a global shortage of rehabilitation services, the majority of physical therapy sessions are conducted at home without professional supervision. This lack of clinical support often leads to low adherence rates and incorrect exercise performance, which can result in longer treatment times, higher healthcare expenses, and poor outcomes for patients. To address this problem, one possible solution is to utilise automated technologies such as cameras and wearable sensors to improve patient access to high-quality rehabilitation services. Despite extensive research on tracking systems for monitoring human movement, many of these solutions are either prohibitively expensive for widespread adoption, or pose usability challenges for patients.

RGB-D sensors - vision cameras combined with a depth sensor and capable of estimating human pose in 3D - have been one of the most popular technologies in this field. However, most of these sensors, like the Kinect, require a GPU for body tracking applications, which increases system costs and reduces the likelihood of large-scale deployment. On the other hand, AI cameras are vision sensors that come equipped with an artificial intelligence chip capable of running complex computer vision algorithms, including pose estimation, within their hardware. They are less expensive than most commercial RGB-D sensors and can operate on a small host computer such as a Raspberry Pi. If they can provide autonomous rehabilitation services, such as motion performance feedback, they would be an ideal solution for large-scale deployment in people's homes.

This thesis presents a proof-of-concept implementation of a rehabilitation system designed for the OpenCV AI Kit with Depth, which is an AI camera with spatial perception capabilities. The system employs a lightweight pose estimation algorithm to track the pose of a human subject as they perform rehabilitation exercises, feeds their joint position data into a neural network, and delivers real-time feedback on their performance via a three-star rating system. The platform demonstrates potential in distinguishing between different rehabilitation exercises and providing realistic assessment scores, but there is room for improvement, such as the creation of a joint-specific feedback mechanism to identify the body parts that contribute the most to each classification.

Contents

2 Introduction

2.1 Background

2.1.1 The Problem: Physical Rehabilitation

Physical rehabilitation is a process in which exercises are performed to correct motor behaviours, with the goal of enhancing physical strength and mobility. According to a study conducted in 2019, approximately 2.41 billion individuals worldwide had conditions that would benefit from rehabilitation - a 63% increase from 1990 [1]. Nevertheless, the World Health Organisation has identified a significant unmet need for essential rehabilitation services, particularly in lower-income countries, owing to factors such as under-prioritisation and inadequate funding from governments, as well as a shortage of rehabilitation professionals and specialised facilities [2].

In the acute stages of illness, in-patient rehabilitation, which involves the supervision of exercises by medical professionals to ensure proper performance, is commonly offered. However, it is neither practical nor economically viable for patients to receive in-person support for every physical therapy session [3]. As a result, it is estimated that over 90% of rehabilitation sessions take place in a home setting [4], where there is no professional supervision, and therefore a lack of performance feedback. Studies have shown that the absence of in-person support from a physical therapist is a significant predictor of non-adherence to home rehabilitation exercises [5].

Non-adherence to prescribed home exercise programmes puts a strain on health services, as it can lead to prolonged treatment times and increased healthcare costs [6], as well as worse treatment outcomes [7]. Furthermore, for those who do adhere, if there is no professional supervision of their motion performance, treatment may be ineffective, or even unsafe, if exercises are performed incorrectly [8]. Moreover, for those who do receive in-patient rehabilitation treatment, the subjectivity of visual perception, the main mechanism for monitoring in behavioural and clinical assessments, can lead to inaccurate or inconsistent advice [9]. Therefore, there is a requirement for a means to monitor patients' physical therapy, including their adherence and exercise performance, without incurring the high costs associated with in-person clinical support for every session, or the inaccuracies related to subjective evaluation.

2.1.2 The Solution: Telerehabilitation

One solution to improve access to quality physical therapy is telerehabilitation: the delivery of rehabilitation services that incorporate information and communication technologies [10]. Telerehabilitation has been implemented both within in-patient facilities, as a supplement to clinical assessment, and in an out-patient environment through the creation of home monitoring services. An optimal solution for out-patient applications would be cost-effective, user-friendly, unobtrusive, and automated. There is a trade-off between these factors for all technologies, depending on the use-case.

One of the most popular technologies for telerehabilitation are marker-less vision-based systems. These systems usually involve the use of a commercial camera, with computer vision methods applied to identify the subject's pose - the position and orientation of their body parts - in captured image frames to determine exercise performance. Marker-less vision systems are relatively low-cost, require no wearable parts, run in real-time, and can be automated. In recent years, RGB colour cameras, which provide 2D image information about pose, have been combined with depth estimation technology to produce RGB-D sensors capable of reliable 3D pose tracking. The most popular commercial example of this is the Microsoft Kinect, for which there has been extensive research on its applicability to rehabilitation [11]. However, the Kinect has some flaws which prevent its adoption as a solution for out-patient exercise monitoring. Firstly, the Microsoft Kinect (Kinect V1) and Kinect for Xbox One (Kinect V2) have been deprecated, leaving the Azure Kinect [12] as the only commercially available option. Secondly, the device is non-configurable, so no custom software can be uploaded into it; as such, the Azure Kinect Body Tracking SDK is the only available method for pose estimation, which requires use of a Graphics Processing Unit (GPU). Furthermore, as well as the cost of the GPU, the device itself retails for $399, which is more expensive than many other commercial RGB-D devices.

Artificial Intelligence (AI) cameras are vision-based sensors with embedded processors for artificial intelligence applications. They are affordable, lightweight, and can run off of a single Raspberry Pi, instead of relying on an external GPU. In addition, these devices tend to be small, lightweight, and much lower in cost than RGB-D sensors - the OpenCV AI Kit with Depth (OAK-D) [13], for example, retails at only $249. If it is possible to implement an exercise monitoring system on an AI camera, and the system can operate in real-time with sufficient accuracy, then it could function as a low-cost and user-friendly telerehabilitation solution. However, there is limited research on the use of AI cameras in a clinical setting compared to other RGB-D sensors.

2.2 Aims and Objectives

The aim of this project is to determine the feasibility of AI cameras as a platform for clinical physical therapy applications, through the development of a telerehabilitation system capable of tracking the pose of a patient and providing them with performance feedback in real-time. The objectives required to meet this aim are as follows:

- Pretrain an image classification neural network architecture on human activity recognition sequences
- Retrain the network to differentiate between rehabilitation exercises and assess general performance
- Implement the system on the OpenCV AI Kit with Depth (OAK-D) [13] AI camera by merging it with an existing pose estimation pipeline

3 Literature Review

3.1 The OpenCV AI Kit with Depth (OAK-D)

The popularity of AI-embedded cameras has increased in recent years, with example use-cases including smartphones [14], transport [15–17], surveillance [18], drones [19], agriculture [20, 21], manufacturing [22] and healthcare [23, 24].

The OAK-D [13] is an AI camera consisting of a 4K-resolution RGB colour camera for visual perception and a stereo pair of 1080p-resolution monocular cameras for depth perception. The OAK-D contains an Intel Movidius Myriad X Vision Processing Unit (VPU) [25], which is capable of running custom neural networks with real-time inference rates. As the device was only made commercially available in 2020, there has been limited research on its efficacy for human pose estimation in comparison to other RGB-D sensors.

A popular application of the OAK-D is for use in aerial vehicles, otherwise known as drones. Due to a requirement for lightweight hardware, most unmanned drones have limited on-board decision-making potential, and therefore have a strong reliance on humans. To solve this problem, [26, 27] proposed a system for autonomous navigation and mapping, in which the OAK-D is attached to a drone via an anti-vibration mount and used for real-time landscape semantic segmentation. An alternative study focused on indoor navigation, using the OAK-D to track the pose of an unmanned vehicle in a warehouse [28]. [29] proposed a method for crowd surveillance in which the OAK-D is mounted to a drone and used to monitor crowd density and ground speed, as well as to identify violent individuals. Moreover, a method for autonomous aerial vehicle racing was suggested in [30], with the OAK-D used to detect ring locations to aid the navigation of the drone.

The OAK-D has also been used for object recognition tasks in robotic systems and internet-of-things devices. Firstly, [31] paired the OAK-D with a mobile manipulation robot, creating an autonomous system for rearranging and discarding items on convenience store shelves. Another study ran a staircase detection and characterisation pipeline on the OAK-D, which helped robots to navigate towards, and align with, staircases in search-and-rescue operations [32]. There have also been applications of the OAK-D in an agricultural setting. [20] implemented the OAK-D within an internet-of-things device to monitor crop growth, whereas [21] paired the OAK-D with a mobile robot to create an autonomous farming application for deployment in wheat fields: the sensor was able to detect diseased crops, and was also used to prevent collision between the robot and the wheat plants.

Finally, there has been some research on the OAK-D as a platform for human tracking applications, such as pose and gesture recognition. Human intention prediction was examined in [22], where the OAK-D was used to identify a subject's head pose, and subsequently forecast their action, with the aim to promote safe human-robot interactions. Furthermore, an OAK-D and Raspberry Pi set-up was used by [23] for the task of fall detection. Despite the use of lightweight hardware, researchers were able to capture data at 26 frames-

per-second (fps). The OAK-D's suitability for human movement analysis was investigated in [24], where its ability to provide feedback on two simple exercises – a side lateral raise and a squat - was assessed. Researchers found that the camera, connected to a Raspberry Pi host computer, could be used to provide feedback on joint angles in real-time, although they noted that for more dynamic motions, greater computational power than a Pi would be required to host the system. They came to this conclusion due to the system's low frame-rate of 7fps with a Raspberry Pi, or 15fps when connected to a laptop. This is likely due to their use of the OpenPose [33] algorithm for human pose estimation, which has a slow run-time on edge devices [34]. This study is the most similar to the one presented in this thesis, and provided a strong basis for further work examining the OAK-D for rehabilitation. We extended their work by implementing a more lightweight pose estimation algorithm on the OAK-D, and testing the system on the same exercises, as well as three others.

3.2 Technologies for Human Pose Estimation

Human pose estimation refers to the process of detecting and locating a person's body in an image or video, in order to determine their position and orientation. It is applicable to many industries, including sports analysis, virtual and augmented reality, robotics, human-computer interaction, and healthcare. There has been extensive research into different technologies for this task, with the most popular being vision-based solutions, comprised of marker-based motion capture systems and marker-less cameras. There are also many solutions involving the use of wearable sensors such as inertial measurement units (IMUs). In addition, some methods enable non-line-of-sight (NLOS) pose estimation, where the subject can be tracked despite the presence of occlusions.

3.2.1 Vision-based Tracking

Marker-based motion capture systems are comprised of multiple calibrated cameras that operate at high frame-rates to track the poses of markers placed on the subject's body. Using time-of-flight and geometric triangulation methods, the system calculates the 3D position for each marker, which can then be used to estimate the subject's kinematics, including the positions and orientations of skeletal joints. There are two main types of marker-based motion capture systems: passive and active. Passive systems detect retro-reflective markers using infrared-emitting cameras, while active systems require the use of infrared-emitting LED markers with cameras to detect their pulses. Active marker systems offer more robust measurements [35], but they require the markers to be wired for power and control, which can hinder the subject's movement [36]. While optical marker-based motion capture systems are considered the gold-standard [37] and can achieve sub-millimetre accuracy [38], they are impractical for most applications. This is attributed to the fixed locations required for cameras, which limits the available motion capture space, and the systems' lack of robustness, with slight changes in camera position requiring the system to

be re-calibrated [39] and marker occlusions leading to tracking failures [40]. Additionally, high costs, lack of portability, and lengthy set-up times restrict the use of these systems to research centres or entertainment studios.

Marker-less vision systems, on the other hand, use computer vision algorithms to identify the subject's shape and movement, without requiring any physical markers. Most marker-less systems employ machine learning algorithms to track the subject's movements from recorded video footage. There are two main choices of light source for these systems: visible light cameras and infrared sensors. A thorough review of human pose estimation methods for visible light sensors is provided in section 3.3.

The most common methods for human pose estimation with infrared cameras are stereo vision, structured light, and time-of-flight. Stereo vision sensors, such as the OAK-D, consist of two monocular infrared cameras, which capture frames of the subject from different angles, and calculate depth by identifying the differences between key-point locations in each image. In structured light methods, a pattern of known geometry is projected onto the scene, and reflected back to the infrared sensor, which uses the pattern's distortion to estimate depth. Time-of-flight sensors illuminate infrared light on the scene, and measure the time taken for the reflected waves to return to the sensor to estimate depth.

Stereo vision is a passive method for depth estimation: it measures infrared waves that is naturally present in a scene. Conversely, structured light and time-of-flight are active technologies, with sensors both transmitting and receiving infrared. Although passive stereo vision tends to produce noisier data, it has an improved operational range in comparison to active methods, and is better suited to outdoor environments [41].

Although infrared sensors are most commonly combined with visible-light cameras for human motion tracking, there have been several studies on the exclusive use of depth maps for pose estimation. [42] created an 'Anchor-to-Joint' algorithm for single depth images, in which a set of anchor points are evenly distributed across the image, and are used to predict the position and depth of joints. Another method named 'DoubleFusion' [43] used monocular depth images to estimate the subject's body shape and underlying kinematic skeleton in real-time. Similarly, [44] regressed skeletal key-points from a 3D point-cloud of a subject to predict their pose.

Although infrared can be used for human pose estimation without a visible light camera, they are most popularly implemented within RGB-D devices - visible light cameras paired with a depth sensor, enabling lightweight 3D tracking. The most popular commercial example is the Microsoft Kinect, which has a custom body tracking algorithm built into its hardware. The first-edition Kinect, known as the Kinect V1, used a structured light approach for depth sensing, whereas the Kinect for Xbox One (Kinect V2) and Azure Kinect contain time-of-flight sensors.

While the Kinect was originally intended as a video game system, its affordable depth imaging technology, paired with accessible pose estimation software kits, has led to its vast popularity as a development platform, with extensive research on its applicability for rehabilitation. [45] implemented a system with the Azure

Kinect which displayed errors of just 3% in comparison to motion capture systems.

One of the biggest challenges in home-based rehabilitation is the lack of patient adherence. Several studies have attempted to solve this by using RGB-D devices to create game-based rehabilitation solutions, which provide a fun and engaging way for patients to complete their prescribed exercises. [46] created a gaming platform for motor rehabilitation, incorporating both single-player and multi-player physical therapy games, with the intention of making prescribed exercises more enjoyable for patients. [47] applied virtual reality gaming to gait rehabilitation, in which the patient takes part in activities such as scoring goals, while their lower limb joint angles and motion trajectories are tracked and sent to a physical therapist. Another study [48] assessed the use of a virtual reality game called 'Stomp Joy' for stroke rehabilitation in a clinical setting, and found that the lower limb function among the experimental group was significantly improved. Furthermore, [49] evaluated a cycling game for elderly patients recovering from hip replacement surgery, in which therapists indicated that the games met the criteria for motor rehabilitation, while patients reported that the system was enjoyable and easy to use. This is evidence that interactive, feedback-oriented rehabilitation can be applied for all age groups.

Despite the widespread use of RGB-D devices in telerehabilitation, their deployment comes with several drawbacks. Firstly, the popular Kinect V1 and V2 sensors are no longer available for purchase. The Azure Kinect [12], which retails for $399, also requires a GPU for body tracking applications, resulting in additional overhead costs. Furthermore, the reliance on a GPU makes the system less mobile and more difficult to deploy as a plug-and-play device, which has led to the majority of Kinect rehabilitation platforms being situated in clinical environments, rather than in people's homes. In comparison, the OAK-D costs $249 and only requires a small host computer, such as a Raspberry Pi. Furthermore, it can run entire data pipelines within its hardware, in parallel with custom neural networks. The sensor's ability to have its software reconfigured by developers gives it a significant advantage for home application settings over other commercial RGB-D sensors, allowing the system to be repurposed as new computer vision algorithms are made available.

3.2.2 Wearable Motion Sensors

Methods based on Inertial Measurement Units (IMUs) involve the combination of a gyroscope, accelerometer, magnetometer and signal transmission chip into a small, low-cost sensor capable of calculating its own 3D orientation and acceleration. Sensors are placed at key points on the subject's body and wirelessly transmit data to a host computer for pose estimation. IMU motion capture has been increasingly applied in a rehabilitation setting, with one study noting that the technology is often used for the repair of human motor function, with a focus on performance assessment [50].

Early solutions for IMU systems required many sensors to be placed on the body for reliable pose estimation. For example, the Xsens MVN [51], a popular

motion capture suit, consists of 17 sensors, while the system in [52] is comprised of 18 plates of IMU and ultrasonic sensors. Although these methods achieved high tracking accuracy, the dense placement of the sensors on the human body is inconvenient and invasive, preventing free movement of the subject, which greatly hinders its applicability to exercise monitoring. As such, there has been research on the estimation of human pose from sparsely-placed IMU sensors. The main challenge of this approach is that the problem becomes harder to constrain as the number of sensors decreases, due to the complexity of full-body human motion.

The first solution to sparse IMU pose estimation was proposed in a model named 'Sparse Inertial Poser' (SIP) [53], which accurately estimated human pose from just six inertial sensors, placed on the legs, arms, waist and head. This was done by applying the statistical SMPL [54] body model, which takes constraints of human kinematics into account, thus making it easier to fit incomplete, ambiguous motion capture data for pose estimation. The main limitation of this method is that it requires data from a full sequence for motion prediction, and therefore cannot run in real-time.

'Deep Inertial Poser' (DIP) [55] extended the work of SIP to create a network capable of mapping IMU orientations and accelerations to body tracking information at 30fps by leveraging both past and future information. The system is able to operate in real-time by using just five frames of future data for each estimation. However, the method struggles with certain poses such as knee bends, which are commonly found in rehabilitation exercises, and positions are tracked with respect to the subject, rather than in global coordinate space.

Subsequently, TransPose [56] extended DIP to provide global translation information, as well as a 90fps frame-rate. They predict the human pose in three stages: firstly, to estimate the positions of five 'leaf' joints from the IMU measurements, then to regress the positions of all 23 joints, and finally to use inverse kinematics to calculate the joint rotations from their IMU positions. To solve the issue of global translation, they assume that the foot in contact with the ground is not moving, and calculate the translation accordingly. Although the system accounted for many of the shortcomings of DIP and SIP, it suffers from the generalisation problem - there are some movements that it is unable to predict, such as those for which the supporting foot is in motion, such as with skating.

As has been demonstrated, IMU technologies for human body tracking are improving rapidly while relying on fewer sensors for the task. However, many have argued that IMU is most beneficial when it is combined with other technologies. For example, the difficulty of providing global translation using IMU sensors alone is more easily addressed using a visual sensor, and IMU sensors are insensitive to occlusion, lighting and human appearance, which are the main drawbacks of a purely vision-based method.

In particular, many studies have looked at the combination of a low-cost RGB-D sensor with an IMU system. A study on upper limbs [57] proved that the fusion of IMU and Kinect data using Kalman filters provided smooth, drift-free tracking, and achieved better results in comparison to using the Kinect alone.

This was further demonstrated in [58], where IMUs were used to compensate for upper-limb joint angle errors detected by the Kinect as subjects took part in movement games.

Although they provide a relatively cost-effective solution for body tracking, without extensive set-up time, IMU-based motion capture systems have disadvantages. Firstly, [50] found that most sensors have a low battery life due to the high power consumption of the signal transmission module. IMUs also suffer from drifts in acceleration sensing results, inaccuracies caused by a slip in its fixed position on the subject, and the sensitivity of the magnetometer to a change in external magnetic field [59], all of which can negatively impact the tracking results. As demonstrated by these studies, the combination of IMU data with artificial intelligence systems can greatly improve its accuracy and efficiency. However, marker-less vision-based systems are frequently chosen over IMU solutions in practice due to the ease of deployment and lower price [60].

Another wearable technology for human pose estimation is ultrasound. A distributed transmitter array broadcasts ultrasonic sound waves, with the time-of-flight calculated by the receiver nodes. [61] and [62] both developed systems of mobile receiver nodes consisting of microphone arrays. Fixed transmitters broadcasted ultrasonic waves, which were picked up by the microphones in each node and used to estimate its complete pose, known as the 6 Degrees-of-Freedom (6-DOF). Both systems achieved centimetre-precision for 3D position. However, this system can incur large biases and high variation, due to the non-linear properties of the transducers used for time-of-flight estimation [63]. Wearable ultrasonic sensors were used in [64] to create a system capable of monitoring foot displacement information during walking to extract gait phases. The system was tested on both healthy and injured subjects, and demonstrated comparable accuracy to a gold-standard motion capture system, with walking speed having no significant impact on results. While these papers have focused on a larger-scale ultrasound set-up, the technology has also been investigated on a smaller scale, such as fine-grained hand motion. For example, [65] created a virtual piano-playing environment in which ultrasound imaging, paired with a transducer fixed on the subject's forearm, was used to estimate finger forces in real-time. Similarly, [66] used ultrasound to image forearm muscles to decode motor intent for hand gesture prediction. There has also been research into the benefits of fusing this technology with other sensors, such as IMUs. One example of this is [67], who used ultrasonic and IMU sensors embedded into shoes to estimate relative foot positions. More recently, [68] combined ultrasonic and IMU sensors to create a real-time 3D hand position and motion tracker. A Kalman filter is used to integrate the 3D acceleration and rotation estimate from the IMU sensor with the 3D position and velocity of two ultrasonic receiver arrays. Although ultrasound is relatively cheap, it still relies on marker placement in order to track a subject, and has a lower accuracy for tracking in comparison to IMUs. In addition, the reliance on microphone sensor arrays may cause privacy concerns if the technology were to be implemented for home-based rehabilitation.

A final wearable sensor for motion analysis is pressure imaging, which involves

the measurement of pressure data to determine characteristics of a subject. One popular application of this technology is for gait rehabilitation, where a patient can wear a shoe or insole to aid with motion tracking. An example of this is in [69], who created smart shoes in which pressure sensors are used to measure the ground contact forces of the subject, while IMUs are placed on the subject's legs to estimate joint rotation. Data from each set of sensors is fused to calculate gait phase and kinematic information, and visual feedback is provided to patients in real-time. Similarly, [70] created a multi-modal system consisting of pressure insoles, IMUs and an RGB-D sensor, to create a virtual coaching system for managing balance disorders. Both of these systems are capable of providing specific gait feedback to patients, but require expensive equipment, with the system designed for treadmill use only. A different application of pressure sensing is seen in [71], where flexible pressure sensors were used to track knee joint postures. The sensor is designed to be worn on the knee-cap, attached to the skin, as a more comfortable solution to IMU monitoring that does not suffer from drift.

3.2.3 Non-line-of-sight (NLOS) Approaches

There are some circumstances in which vision or wearable sensors, despite their low cost, are not suitable for tracking applications. For example, the presence of RGB cameras in intimate locations such as bedrooms, bathrooms and hospital wards would likely cause privacy concerns. Furthermore, full-time wearable sensing solutions are often regarded as invasive. Unobtrusive pose estimation sensors are necessary for applications requiring placement within homes, hospitals, or elderly care centres, with example use-cases including for fall or seizure detection. Some technologies being researched in this field include radar, LiDAR, and WiFi.

Radar systems emit high-frequency electromagnetic signals, which are reflected back to the receiver by target objects The frequency of the returning signal, and its angle of incidence, are used to estimate the speed and direction of the target object. The application of radar with deep learning for human motion recognition was assessed in [72], who argue that the main advantage of radar for indoor monitoring is the preservation of monitored individuals' privacy, thus enabling the tracking of their motion in locations where people may be uncomfortable with the presence of vision-based sensors. The ability of radar to identify velocity components of moving body parts is demonstrated in [73]: researchers use time-frequency signal analysis, alongside a machine learning approach, to classify human walking motions. The notion of unobtrusive tracking is further explored in [74], who created a gait analysis system capable of identifying pathological or assisted walking patterns in test subjects by measuring changes in radar micro-Doppler signatures. Although these papers focus on general gait applications, radar can also be used for specific joint tracking, such as in [75], where millimetre-wave radar is used to track and classify the subject's hip joint motions in physical therapy exercises. Another particular benefit of radar is that it can track internal organ movements, due to its ability to penetrate opaque

objects. An example of this is shown in [76], who created a system capable of respiration monitoring for human subjects. The main advantages of this application is that it does not require the subjects to wear any sensors, nor are they required to remain static for accurate measurements to be taken. Although radar is an useful technology for unobtrusive tracking and identification of internal movements, it is unlikely to replace RGB-D sensors for human pose estimation. One study [77] used millimetre-wave radar to predict human poses, with the Kinect used as the ground-truth, and found that the system could accurately identify the location of the subject, but struggled to detect granular joints such as the hands. Therefore, for the task of full-body pose estimation, such as in our take-home telerehabilitation system, vision-based sensors are a more suitable choice.

Another tool that has been explored for human pose tracking is light detection and ranging technology (LiDAR), which measures the absolute distance of objects based on the time-of-flight of emitted lasers. Similarly to radar, LiDAR is non-intrusive, but has the added benefit of being more portable and cost-effective, with a trade-off of sensitivity to the colour of objects, and a reduction in accuracy at greater distances from the sensor. Researchers in [78] used a 2D LiDAR placed on the ground to acquire gait data from subjects' ankles. Although the LiDAR could only track ankle positions in the transverse plane, the accuracy of detected gait parameters was comparable to a gold-standard motion capture system, although it struggled to make detections when the subject was over 10m from the sensor. Another implementation of 2D LiDAR for gait tracking [79] placed the sensor at shin height, with the aim to track the positions of human legs. Their method is also capable of tracking walker users by identifying and removing data representing walker legs. While 2D LiDAR has shown promising results for low-cost gait analysis, similarly to other technologies, it is most effective when used in a multi-modal system. For example, [80] used a combination of a 3D LiDAR and 10 IMUs to track a subject's position and pose in a small office space. Although the system was more accurate with LiDAR than with the IMUs alone, the use of 10 IMU sensors, as well as a need to perform calibration, makes the system non-ideal for simple, home-based rehabilitation applications where a professional is not present to set up the system every time it is used. Similarly, [81] used two LiDARs and 10 IMU sensors to estimate human skeleton parameters. This was done by extracting point-cloud information from the LiDARs, and capturing the orientation of body parts from IMUs, after which the data is fused to estimate joint positions and correct IMU drift. The researchers argued that their system is comparable to the state-of-the-art vision motion capture systems, without a dependence on markers and with little set-up time. Therefore, this multi-modal system is more beneficial as a low-cost replacement to a marker-based motion capture system, rather than as a solution for home-based monitoring.

WiFi is advantageous for human pose estimation because of its prevalence and ability to circumvent occlusions similarly to radar and LiDAR. A key benefit of many modern WiFi-enabled systems is their ability to operate using existing WiFi infrastructure. The first application of 3D body tracking of moving

subjects was seen in Wi-Mose [82], which uses deep learning to extract pose and position information from the WiFi's Channel State Information (CSI) - the ratio between the transmitted and received signal wave. Another implementation is found in Winect [83], which monitors the angle of arrival of WiFi signals reflected from the human body, using this information to estimate the trajectories of moving limbs before constructing the subject's 3D skeleton. Another algorithm named DensePose [84] maps the phase and amplitude of WiFi signals to coordinates of the human body surface by converting CSI signals into 2D feature maps, which are fed into a neural network to estimate pose. However, the system fails when subjects perform actions that are rare within the training set, as well as when there are many subjects in one capture. Similarly, GoPose [85] monitors the poses of moving subjects by extracting the angle of arrival of incident WiFi signals, which represent spatial information of different body parts, and subsequently translates this into joint locations in physical space. GoPose utilises two different deep learning models - CNNs for capturing spatial features of body parts, and LSTMs for estimating temporal motion features - and unlike DensePose, was able to identify poses that were unseen within the training data. Developments in WiFi for body tracking have led to its implementation for rehabilitation: one example of this is Wi-PT [86], which uses existing WiFi infrastructure to monitor patients during their home-based rehabilitation programs. Wi-PT was shown to recognise different scales of movements, from hand and wrist motion to whole-body equipment-based exercises, at a high level of accuracy, and could also identify which person was performing each exercise. This paper demonstrates that WiFi is an ideal solution for long-term exercise monitoring, even in a clinical application. However, there is no research so far that has been able to provide specific performance feedback based on WiFi-tracked human poses, so it is likely better suited as an alternative to radar and LiDAR for applications such as fall detection.

3.3 Vision-based Human Pose Estimation Algorithms

Monocular vision-based methods require the estimation of human pose from a single view-point, as either a colour frame alone, or with corresponding depth information. Solutions to this task differ greatly with respect to data input modalities, joint prediction methodologies, and output representations.

Two popular approaches for human pose estimation are top-down and bottom-up methods. The top-down approach precedes pose estimation with a human detection step, which produces a bounding box around each person within a scene and subsequently identifies the body parts. One example of a top-down approach is AlphaPose [87], which directly regresses key-point locations within a bounding box using deep learning. In comparison, Mask R-CNN [88] uses an image segmentation mask to identify individual body parts within the box. While regression-based methods can be more accurate due to the direct prediction of key-points as opposed to image segments, detection-based methods are often more robust to occlusion and clutter. On the other hand, bottom-up approaches focus on the identification of body parts within a scene

before grouping them to form individual poses. An example of this approach is OpenPose [89], which uses key-point identification to locate key-points before implementing a graph-based method to group them into skeleton topologies, such that the positions and orientations of each joint connection are encoded as 2D vector fields known as Part Affinity Fields. While top-down approaches benefit from greater accuracy, due to the higher resolution of each image passed to the pose estimator, the run-time increases drastically for each additional person within a scene; as such, bottom-up approaches tend to be preferable for use-cases involving large groups of people.

One common output representation for human pose estimation is a heat-map which encodes the likelihood of each pixel within an image belonging to a given body part. One example of a use of this method is TransPose [90], which regresses key-point locations for each joint from its corresponding heat-map. Although this method is robust to occlusions and often results in greater accuracy than direct regression-based methods, heat-maps can be computationally expensive to develop, and as such, this method is likely unsuitable for scenes containing many people. In contrast, TFPose [91] uses transformers to directly identify each body joint from the input image, in a way that naturally exploits the structural dependencies between joints. For offline processing, heat-map regression is favourable due to the greater accuracy and robustness, however for real-time applications it is likely infeasible, and regression-based methods are preferred.

3.3.1 Lightweight Methods

Although most research on human pose estimation has focused on achieving the highest possible accuracy and robustness to occlusion, many applications, such as the one in this thesis, require a lightweight method that can run in real-time on edge devices. There has been some research into finding such methods that can run at high frame-rates without a great accuracy trade-off.

Firstly, the creators of OpenPose implemented an optimised, lightweight version [92] that can run at 28fps on a CPU, with just 15% of the complexity of the original method, while maintaining a similar quality. This was done by parallelising key-point extractions, upsampling the input image, and removing extra memory allocations. They found that the ratio of accuracy to network complexity was increased by a factor of 6.5 when a MobileNet feature extractor was used. This paper demonstrates the ability to make pose estimation algorithms compatible with edge devices by simply increasing the efficiency of resources used.

Another lightweight method is known as SimpleBaseline [93], which uses heat-maps and a ResNet backbone with a bottleneck to produce state-of-the-art pose tracking. The method uses a human detector in conjunction with an optical flow tracker to provide temporal information about the subject's pose. The use of previous frames enables the system to be robust to occlusion and motion blur, as it uses examples of prior high-confidence person detection to aid the pose estimation of the current frame. The bottleneck block was subsequently

used by the creators of the Lightweight Pose Network [94] to achieve similar performance results to SimpleBaseline with only 9% of the model size, and 11% of its complexity. They achieved 17fps on a non-GPU platform with a ResNet-50 backbone, which is a sufficient rate for real-time use-cases.

In contrast to the examples above, which prioritise multi-person pose estimation, the creators of [95] used a similar architecture to SimpleBaseline to implement a fast and lightweight network for single-person use-cases. They created a lightweight bottleneck which uses structural similarity measurements to decrease the model size, as well as an attention mechanism for modelling contextual information. The bottleneck drastically reduces the parameters and floating-point operations, allowing for the architecture to be deployed on limited-resource CPUs.

Finally, BlazePose [34] is an ultra-fast, lightweight, single-person pose estimation algorithm for edge devices. An encoder-decoder network is used to predict heat-maps for all joints, which are then fed into another encoder to directly regress the coordinates. During inference, the heat-map branch can be discarded to enable the model to run in real-time on mobile phones. They use a top-down approach, with a human detector that only runs when the tracker cannot identify a subject. The researchers observed that the strongest indicator to a neural network about a subject's torso position is the face; as a result of this, the network relies on the assumption that the face is visible in all frames, and uses their BlazeFace [96] detection algorithm as a proxy for person detection. The assumption of face visibility is suitable for our use-case, which will require the subject's entire body to be in frame for performance assessment. According to [34], BlazePose Full (with heat-maps) was able to outperform the heavier OpenPose algorithm on yoga and fitness cases, while running between 25 and 75 faster on a single mid-tier phone CPU, in comparison to OpenPose on a 20-core desktop CPU. However, another study [97] found that although BlazePose had a far greater processing speed, and a correlation of 0.8 with OpenPose estimations, they identified a decreased quality with its prediction of key-points and determined that the algorithm is not yet clinically viable. However, the data used for this study was comprised of RGB smartphone videos, and z-position information was discarded. It is likely that 3D outputs would greatly benefit pose estimation, especially if the model is provided with additional depth information about the scene, as with our application. In addition, its ultra-fast run time, while maintaining a high performance quality, makes this algorithm the most suitable for use with edge devices.

3.3.2 Activity Recognition and Assessment

In telerehabilitation, it is especially important for the system to not only display or record motions, but to also provide feedback to the user, in order to prevent incorrect motions from hindering recovery, as well as to encourage patients to engage with their rehabilitation.

[98] explored different metrics for automated performance evaluation of physical therapy exercises. The paper explained that in order to determine

'correctness', sequences can be assessed using quantitative methods, which they divided into two categories. Model-less metrics assess motions based on raw time-series joint trajectories, whereas model-based metrics calculate consistency by modelling motion sequences through a set of latent states describing the statistical distribution of motion dynamics. Each sequence can then be assigned a score based on quantitative analysis, and marked as 'correct' if the score exceeds a threshold.

Most existing methods for exercise performance take a classification approach, in which the model deems the exercise as either 'correct' or 'incorrect'. This approach does not have the capacity to detect varying levels of movement quality or identify incremental changes in patient performance over the duration of the rehabilitation program. For maximum utility, telerehabilitation systems should go beyond classification and output feedback on the motion, including advice to the patient on how they can improve their score, in a way similar to a physiotherapist or clinician.

An example of a model-based application is seen in [99], who proposed a log-likelihood metric to train a deep learning network to assess Vicon rehabilitation data from the UI-PRMD dataset [100]. They argued that model-based statistical methods are able to encode inherent randomness in human motion, making them more robust to spatial and temporal variations in data. Their algorithm outperformed model-less distance-based metrics for most of the actions, and asserted the applicability of deep learning to the motion assessment problem. Contrastingly, [101] used two Kinect V2 sensors to measure model-less metrics such as angles and distances between joints, in order to evaluate stroke patients and healthy subjects according to the standard stroke rehabilitation scale – the Fugl-Meyer Assessment [102]. Their network achieved near-perfect success with score classification, demonstrating the feasibility of using low-cost RGB-D sensors for a home-based rehabilitation assessment platform.

3.4 Datasets

3.4.1 Human Action Recognition

The first task of the project was to choose a human motion capture dataset for training the machine learning algorithm to perform HAR. Requirements of the dataset include capturing pose in 3D, providing kinematic-based pose tracking coordinates and orientations (as opposed to a volumetric model), having a sufficient amount of data for each activity, an extensive range of actions, many subjects, and that it is publicly available for download. Many of the popular datasets for human pose estimation, such as DensePose-COCO [103] and MPII [104], use RGB image data only, with 3D positions estimated from 2D information only. Since the AI camera has depth estimation functionality, we focused exclusively on datasets that track human pose in 3D.

There are some 3D motion capture datasets which are particularly popular within the HAR literature for benchmarking algorithms. Firstly, Human 3.6M [105] contains 3.6 million poses of actors performing everyday motions such as

taking photographs, greeting, and eating. The dataset provides RGB, depth (time-of-flight) and 3D body scans of all subjects. The main drawback is that tracking data is not synthesised into kinematic joint position estimates. Furthermore, KIT [106] is a large-scale database collected using the Vicon C3D; it contains a vast assortment of human motions, including those involving objects and multiple subjects. However, body tracking information is augmented using a volumetric pose representation, rather than kinematic joint position estimations; this would require an additional step of converting their data into a skeletal format compatible with the AI camera's pose detection.

In comparison, the MoVi dataset [107] contains 21 activity sequences for each of the 90 subjects, and provides coordinate predictions for 20 tracked skeletal joints for each frame. It also uses a Vicon motion capture system, operating at 120 frames-per-second. The large number of subjects, wide range of actions, and use of a kinematic skeleton model, make this dataset suitable for pre-training in our application.

3.4.2 Telerehabilitation

For later project stages, we have considered different body tracking datasets specifically centred around rehabilitation exercise applications.

Firstly, the Toronto Rehab Stroke Pose Dataset [108] contains recorded upper-limb rehabilitation exercises for stroke patients, captured using a Kinect V2. The main drawback of this dataset is that exercises were centred around the use of a robotic arm; for take-home telerehabilitation, it is a priority to have the system be as low-cost and simple to use as possible, and as such, reliance on an expensive and complex piece of equipment makes this dataset unsuitable for our application.

Secondly, the Kinect 3D Action dataset [109] contains Kinect V2 body tracking data of 13 actions that represent common clinical assessments of balance, mobility, and physical performance. One of the benefits of this dataset is that there are 54 subjects, with an age range between 18 and 81, and a large standard deviation of ages. This means that the dataset contains large motion variations between skeleton poses for a given action, which is representative of real-world applications. Although no 'correctness' label is provided, the inter-class variation for each action would help with training an assessment algorithm to differentiate between performances of the motion.

IntelliRehabDS [110] collected data using a Kinect V2 from 15 real rehabilitation patients as well as 14 healthy subjects, each performing repetitions of 9 gestures. Each sequence also has a 'correctness' label. Although the majority of sequences are marked as 'correct', the researchers have provided software to synthesise 'incorrect' motions for the generation of training data. The use of real rehabilitation patients is beneficial, however the lack of real 'incorrect' data was a downside.

Another dataset considered was UI-PRMD [100], which simultaneously combines a Vicon system with a Kinect sensor to record data from ten healthy subjects, each performing 10 rehabilitation exercises. The large number of sequences per subject (10 each for 'correct' and 'incorrect' movements), the use of a motion

capture system for data collection, as well as the large number of both correct and incorrect motions for each exercise, made this dataset a good fit for our application. However, when the dataset was downloaded, there appeared to be a lot of missing information in the Kinect data, likely due to reflectivity issues caused by motion capture markers [111]. In addition, the Vicon data was not processed with any skeletal rendering, meaning that only the coordinates of the surface-level markers themselves were provided. As the application we are constructing relies on estimated skeletal pose joints, the UI-PRMD was a non-ideal dataset for our application.

Finally, the KIMORE dataset [112] uses a Kinect V2 to collect data for five rehabilitation exercises specific for lower back pain. The V2's reliability for use in this setting was validated by analysing its accuracy with respect to a gold-standard system prior to data collection, with the results indicating that it is a suitable tool for dynamic postural assessment [113]. There are 78 subjects, of which 44 are healthy and 34 suffer from motor dysfunctions due to either stroke, Parkinson's or back pain. Additionally, each sequence is assessed by clinicians, who filled out a questionnaire containing questions related to both the primary objectives of the exercise and the postural performance of the subject. The Primary Outcome and Control Factor (postural) scores were summed to produce each sequence's Total Score, and each of these scores are provided in the dataset. The provision of scores is a key benefit of this dataset for use in performance assessment, as opposed to other datasets which either provide only a correctness label, or no label at all. In addition, KIMORE's suitability for real-time rehabilitation monitoring was validated prior to the dataset's release [114]; the Kinect V2 was shown to be sufficiently accurate for dynamic postural assessment, and the data collected was demonstrated to be suitable for real-time applications. We decided to use this dataset in our application due to its large number of subjects, its incorporation of real patients, the provision of performance scores, and its use of simple rehabilitation exercises.

4 Methodology

4.1 Human Action Recognition

4.1.1 Transfer Learning with ResNet-50

To determine the feasibility of our novel machine learning solution, we opted to implement it on a popular, well-tested architecture that is capable of being applied to solve many problems. In this instance, we decided to use ResNet-50 [115], a 50-layer neural network formed by stacking convolutional layers with residual connections. This architecture was trained on the ImageNet [116] database of over 1 million labelled images, and is widely regarded as a state-of-the-art solution for image classification.

Joint position data is often provided as 3D coordinates for each joint within the topology of a skeleton model. Therefore, for a body tracking sequence, the data has dimensions (frame index, joint position index, coordinates). This is similar to the shape of an image: (height, width, channels), where the 'channel' dimension represents the pixel intensity of each of Red, Green and Blue. The similarity of dimensions between pose information and images enabled us to convert data from action recognition sequences into an RGB input format; consequently, the ResNet-50 architecture would be able to classify the activities within each sequence similarly to how it classifies objects within images.

One problem with this form of transfer learning is that the images produced from skeletal sequences differ substantially to the images that ResNet-50 was trained on, as they contain no physical objects. This means that the ResNet-50 layers would need to be re-trained for this application. Although this could have been done with a rehabilitation dataset directly, captured with an sensor similar to the AI camera, these datasets are fairly small, and the lower accuracy of RGB-D sensors may result in a lower accuracy of the rehabilitation network, especially when applied in practice. Therefore, we decided to pre-train the ResNet-50 network, with imported ImageNet weights, on a large dataset of human activity sequences, captured at a high frame-rate from a motion capture system.

4.1.2 The MoVi dataset

The chosen dataset was MoVi [107] - a large, multi-purpose motion and video dataset of 90 subjects - 60 female and 30 male - performing a collection of 21 action and sports movements. Sequences are performed five times by each subject, each with a different number of motion capture and IMU sensors to collect pose information. We chose to use the 'F' sequence set, in which subjects wear 67 motion capture markers, and for which there is no IMU tracking, as this data was not relevant for our application.

The MoVi research team augmented data with both MoSh++ [] and Visual3D [117] to compute the skeleton. MoSh++ estimates the body shape, pose, and soft tissue deformation, and provides this information as joint angles. The MoVi authors recommend this software model for animation applications. Visual3D,

on the other hand, is a biomechanics analysis software that calculates joint positions as both orientations and absolute 3D positions. It is recommended for medical applications such as gait analysis and is deemed to be more accurate by the authors of the dataset paper. As most RGB-D sensors use joint positions, rather than angles, and there was a desire to prioritise accuracy in our application, we opted for Visual3D skeletal rendering.

The 'F' dataset is provided as a collection of MAT files, with one file for each subject. The file contains markers which indicate when each sequence begins and ends, as well as the name of each sequence in chronological order - this order is different between subjects. In addition, of the 21 action sequences, one is considered a 'random motion' in which subjects perform a spontaneous action at their own discretion. Since this action was different for each subject, we omitted it from training. 3D coordinates consist of the 67 motion capture markers, as well as 20 Visual3D 'virtual markers' representing joint position estimates.

4.1.3 Pre-Processing

It was decided that for pre-training, 120-frame samples, representing one second of motion capture data, would be taken from the centre of each sequence. In addition, only the virtual marker positions were used, as these align closely with other skeletal topologies, such as BlazePose, whereas the motion capture markers are surface-level. In some instances within these samples, positional information was missing, and coordinates were set to the origin. This is because the Visual3D software does not attempt to estimate joint locations with the presence of occlusions. Although this means that the joint positions provided are highly accurate, it also required us to manually deal with outliers before converting the sequences to images.

Figure 1 shows an outlier mask generated for a 120-frame sample of the 'vertical jumping' sequence, performed by Subject 1. The joint indexes, from 0 to 19, are shown on the x-axis, whereas the frame indexes are represented on the y-axis. The outlier mask graph indicates that joints 8 and 16 have missing information for a small number of frames. Given that the entire sample represents one second of motion, due to the motion capture system operating at 120fps, these gaps are acceptably small for a linear interpolation method to be applied. Therefore, positional data of these joints for non-outlier frames was used to fill in the gaps.

A visualisation of the impact of this outlier removal technique is shown in Figure 2. Since outlier joints are set to the origin, as shown in Figure 2(a), the pixel intensities of each of the red, green and blue channels for the RGB representation are set to 0, corresponding to black lines on the image, as shown in Figure 2(c). Since the missing frames represent a fraction of a second, linear interpolation provides a reasonable estimate of the true joint positions for missing data, as demonstrated in 2(b). The 'fixed' RGB image, which was used as an input to ResNet-50, is shown in 2(d).

This method of outlier removal was applied to all data samples used for pre-

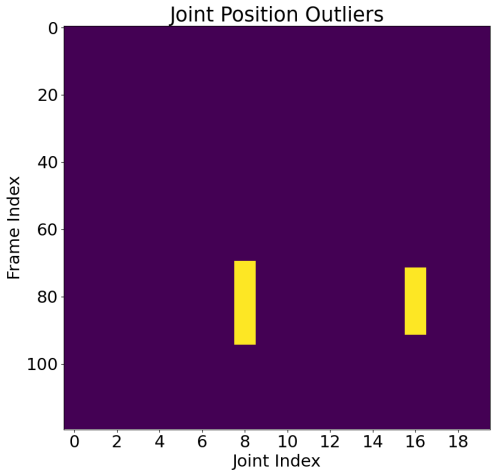


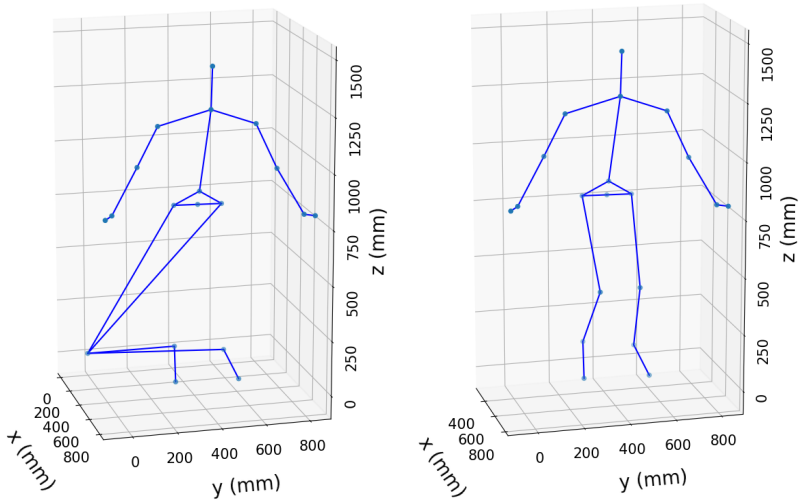
Figure 1: An outlier mask of a 120-frame sample from the MoVi dataset of the 'vertical jumping' action, performed by Subject 1. Outlier locations are shown in yellow, where the x-axis indicates the joint index, and the y-axis represents the affected frames. In this case, joints 8 and 16 - the knee joints - are impacted.

training. After linear interpolation, sequences were reshaped from dimensions (120, 20, 3), representing frame count, joint position index, and coordinates respectively, into the ResNet-50 input shape of (224, 224, 3), resulting in RGB images such as the one in 2(d).

4.1.4 ResNet-50 Pre-training

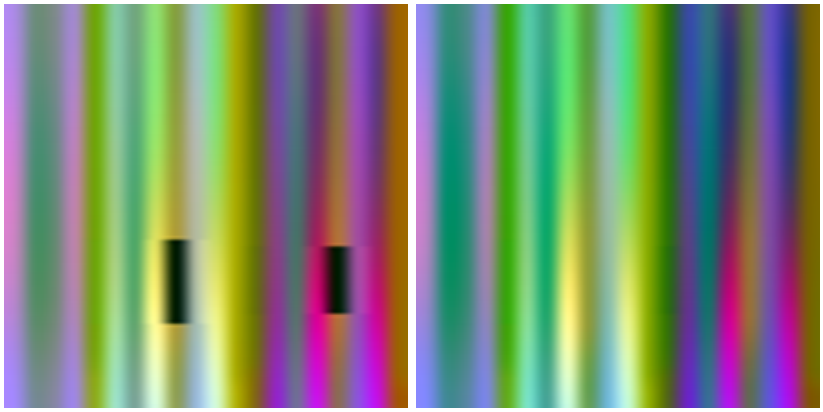
Tensorflow and Keras were used as the platforms for machine learning within this project. ResNet-50, with its pre-trained ImageNet weights, was loaded into the program and the top layers, originally used for animal classification, were removed and replaced with Dense layers of lengths 40 and 20 respectively. The 20-neuron layer was the network's output, representing each of the 20 actions.

The network trained with a batch size of 8, and a learning rate of 0.001, with a checkpoint callback set up to save the best model over 400 epochs of training. A validation split of 0.2 was also used. Samples were saved in directories corresponding to their action class, and data was loaded into the network using the Keras function *image_dataset_from_directory()*, which generated one-hot encoded vectors for each sample corresponding its directory name.



(a) Skeleton with outliers

(b) Skeleton after linear interpolation



(c) RGB image with outliers

(d) RGB image after linear interpolation

Figure 2: A 120-frame sample from the MoVi dataset of the vertical jumping action for Subject 1, with and without outliers. The missing knee joint data, shown in (a), correspond to the black columns in (c), which align with the outlier mask from Figure 1. Linear interpolation was applied to estimate the joint positions for missing data, with results shown in Figures (b) and (d).

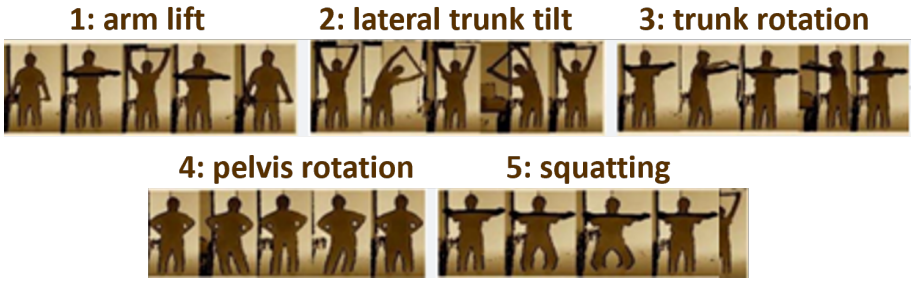


Figure 3: The five exercises in the KIMORE dataset. This image was adapted from the one included in the dataset paper [112].

4.2 Rehabilitation Assessment

4.2.1 The KIMORE dataset

The KIMORE dataset contains data of 78 subjects performing five common physical therapy exercises for lower back pain, shown in Figure 3. There are 44 healthy subjects, with no history of neurological or musculoskeletal problems, and 34 subjects suffering from chronic motor disabilities. Furthermore, the researchers split the subjects into five categories; of the 44 healthy subjects, 12 were physiotherapists and experts in back pain and postural rehabilitation (group E), and the remaining 32 were non-experts (NE). In addition, of the 34 subjects with pain and postural disorders, 10 had stroke (S), 16 had Parkinson's disease (P), and 8 had back pain due to spondylosis (B). Furthermore, the subjects exhibit a wide spread of ages between them, and an even distribution of gender. These factors all contribute to the KIMORE dataset's ability to generalise well - many other datasets contain only healthy subjects, or a small number of participants within a narrow age range. It is hugely beneficial to train a network for rehabilitation assessment on real patients who would use the application once it is available.

The five exercise sequences were each performed with five repetitions, with the subject stood 3m from the Kinect V2 sensor. Body tracking information is provided both as joint positions and joint orientations, and depth videos are also included for most of the subjects. There is also supplementary information about each subject, detailing their age, gender, group, and clinical scores.

Arguably the greatest benefit of this dataset over alternatives is the provision of a clinical assessment for each exercise sequence. The KIMORE researchers created the Exercise Accuracy Assessment Questionnaire [118] as a mechanism for assessing exercise performance quality, shown in Figure 4. The first three questions investigate the primary objectives of each exercise, whereas the rest are concerned with the correct posture of seven body segments. The Primary Outcome and Control Factor scores are provided separately for each sequence, alongside the Total Score, which indicates the overall quality of the exercises as the sum of the ten question scores. A box-plot of the distribution of Total

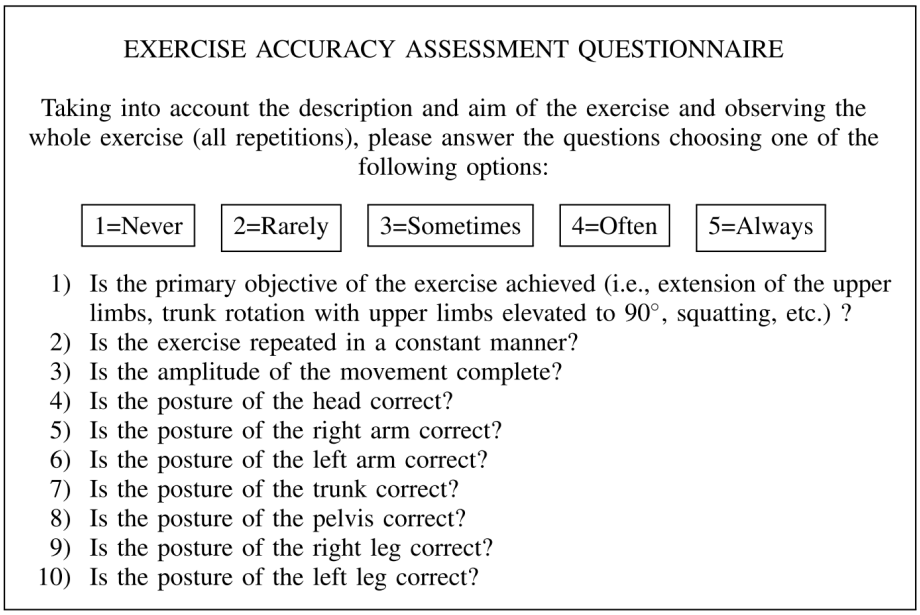


Figure 4: The Exercise Accuracy Assessment Questionnaire, developed in [118], used to provide the clinical scores in the KIMORE dataset. This image is from the KIMORE dataset paper [112].

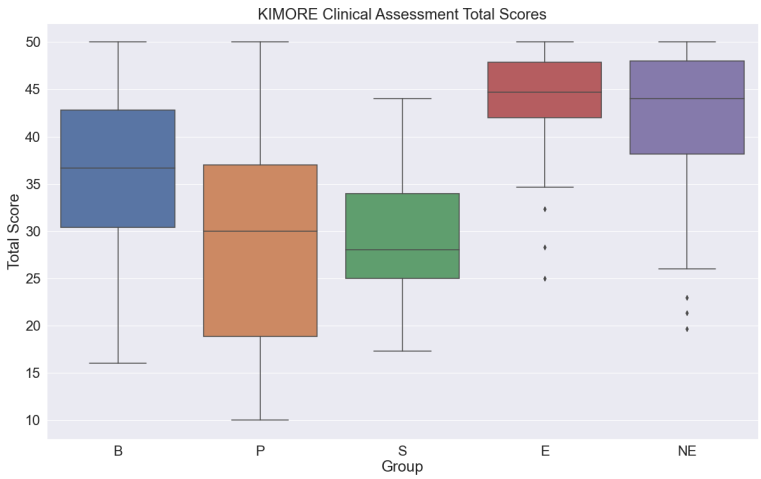


Figure 5: A box-plot of the spread of Total Scores for each group in the KIMORE dataset. As shown, Experts (E) and Non-Experts (NE) tend to achieve higher scores, whereas patients are more likely to achieve lower scores, as expected.

Scores for each group is presented in Figure 5, demonstrating that Experts and Non-Experts were more likely to achieve higher scores than patients, as would be expected. However, it also shows that the subject's group is not necessarily a determining factor of their score - there were low and high scores performed by all subjects, regardless of their expertise and physical health.

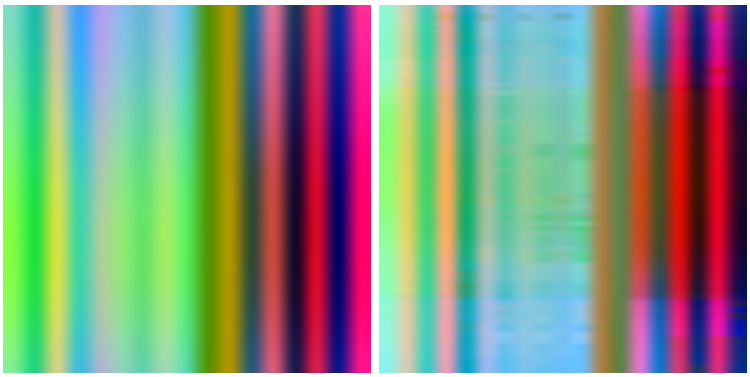
4.2.2 Pre-Processing

Although the Total Score of an exercise is provided as a mark out of 50, we did not use the score directly in our application. The first reason for this is that the scores are subjective, as they were determined by a clinician from visual inspection, so assessments within a few points of one another likely exhibit similar overall performance. In addition, the telerehabilitation network is being developed for real-time applications, so it was necessary to take small samples of each exercise for analysis, rather than the whole sequence. This poses a problem of consistency, since a subject's performance may fluctuate throughout the duration of a sequence. Finally, even if the network is able to learn to regress scores for small samples of data, the accuracy is predicted to drop considerably when the system is implemented on the AI camera, due to network inputs coming from a different sensor and pose estimation algorithm that the network is not trained on. Therefore, we abstracted the scores into a 3-star rating system, such that the highest scores were given a 3* rating, whereas the lowest scores were awarded 1*.

Figure 6: A histogram plot of the distribution of Total Scores in the KIMORE dataset for all sequences. As shown, scores are far more likely to be awarded a high score than a low score. The class thresholds used to abstract scores into each rating category are shown in red.

In our initial implementation of this, we split the scores such that there was an even number of sequences in each rating category. This led to a system in which the bottom tertile of sequences, awarded a 1* rating, contained a score between 0 and 37.5, while the top tertile consisted of scores above 44. This was due to the uneven distribution of sequence scores in the dataset, and resulted in thresholds between rating categories that were close together. In practice, a score of 37.5 may represent a well-performed sequence with minor errors, whereas a score of 20 may represent an unsafe performance that could be detrimental to the patient's recovery, therefore it makes little practical sense to group such scores together. In addition, the tight class boundaries make it difficult for the network to differentiate between rating classifications, due to the variation of a subject's performance within a sequence, as well as the subjectivity of clinical assessment.

Instead, we opted to split the rating categories roughly into tertiles out of 50, irrespective of the number of sequences in each class, as shown in Figure 6. As a result of this method, over half of the sequences were awarded 3* ratings, whereas only the poorest 15% of scores were given a 1* rating. Class weighting, through provision of a cost matrix, was then used to account for the uneven distribution, such that the network penalised incorrect classification of 1* sequences more heavily than the other rating classes. This method is beneficial for our application, which aims to provide telerehabilitation to prevent poor exercise performances, which are most likely to exacerbate injuries in



(a) MoVi (Subject: 33) (b) KIMORE (Subject: Expert 7)

Figure 7: A comparison of a squatting motion sample from both the MoVi and KIMORE datasets, demonstrating the difference in coordinate jitter and the impact it has on the quality of RGB images used for training.

practice. The differentiation between good performances with minor flaws, and perfect performances, is not within the scope of this application, and may be an area for future work.

A sliding window was used to extract multiple samples from each sequence for training, rather than just one central sample as was done with MoVi. 90-frame samples, with a 45-frame overlap, were taken from each sequence, representing 3 seconds of sequence data (the KIMORE dataset was captured with a Kinect V2, which operates at 30fps). This was deemed to be an appropriately long time period to be able to recognise the exercise and its quality, as well as being short enough for real-time feedback application. Samples of the first and last 90 frames of each sequence were omitted, to prevent the use of data in which the subject was not constantly performing the exercise.

Although the MoVi and KIMORE datasets do not contain the same activities, some of the 'random motion' sequences in MoVi contain the squatting action, which is one of the exercises performed in the KIMORE dataset. Normalised[1] RGB image representations of a MoVi and KIMORE squatting sequence are compared in Figure 7, for which the KIMORE sequence was awarded a 3* rating. As is shown, the images generated from KIMORE data contain significantly more jitter, due to the reduced accuracy of the Kinect V2 for motion capture in comparison to the gold-standard Vicon. Although the use of noisy input images may hinder classification accuracy, it is more representative of the data quality that the OAK-D would produce in practice; therefore, it would not be beneficial for our application to train the model on perfectly smooth data, as the system would not be robust to jitter in the pose estimated by the BlazePose algorithm.

[1]For more information about normalisation, see section 4.3.2

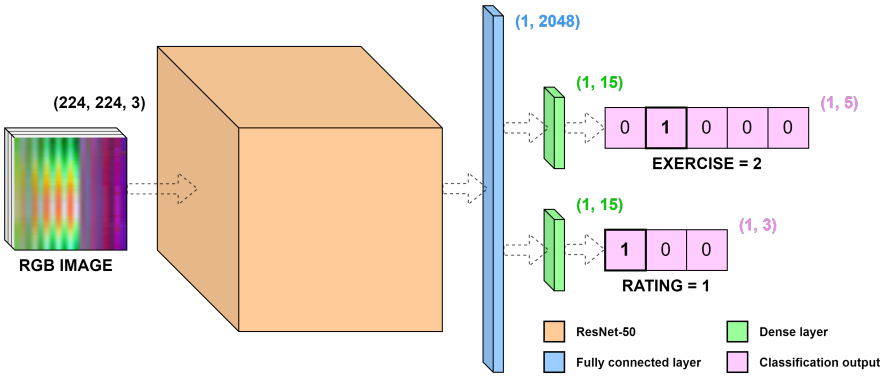


Figure 8: A high-level representation of the rehabilitation neural network, showing the dimensions of data at each point.

4.2.3 Training

The telerehabilitation neural network was designed similarly to the one used for pre-training: the network took an RGB image as input, and passed it through ResNet-50, followed by Dense layers for generating the classification. The ResNet-50 layers of the MoVi model, pre-trained for human activity recognition, were transferred to the new model, while the rest of the network was discarded. Instead, the ResNet-50 output was fed directly into branches for exercise and rating classification. The reason for this is that the output layer of ResNet-50 is a 2048-neuron feature vector, representing a high-dimensional embedding of the input image pose; the use of a lower-dimension Dense layer as a direct output of ResNet-50 would reduce the amount of information available to each classification sub-network. Since the neurons within the exercise and rating branches have their weights optimised for different tasks, it was beneficial to use separate Dense layers to reduce the dimensionality down, which are optimised for solving that specific problem.

A diagram of the network is shown in Figure 8. The input layer matches the dimensions of the RGB image, which is followed by the imported ResNet-50 layers, including the fully-connected feature vector, which acts as input to both of the Dense layer branches for exercise and rating classification respectively. Outputs are provided as probability vectors, such that the maximum-likelihood classification from the network is equal to the vector index with the largest value.

The network was trained for 400 epochs with a learning rate of 0.00001, a batch size of 8, and a checkpoint callback monitoring rating prediction accuracy. It was noticed that the exercise classification branch was prone to overfitting, so an early stopping mechanism for the exercise loss was implemented to prevent this.

4.3 AI Camera Integration

4.3.1 Overview of OAK-D and Depth-AI

The OAK-D is an AI camera capable of spatial perception. It calculates depth using stereo vision: a pair of cameras take captures of the scene, and the disparities between key-points common to both sets of frames are used to estimate depth. This information is paired with colour frames from a 4K-resolution camera to provide information about the scene in three-dimensions. The main difference between the OAK-D and standard RGB-D sensors on the market, such as the Kinect, is that the OAK-D contains a Visual Processing Unit (VPU) - a powerful processor that is capable of running custom neural networks for both visual and depth perception in real-time.

Models must be converted to Binary Large Object (BLOB) file format before being uploaded to a VPU device. This was done using an online Blob Converter tool, which first converts the model to an intermediate OpenVINO representation, before recompiling it in the correct format. However, many processing steps were required to convert the rehabilitation neural network, which was saved as a Tensorflow SavedModel, to a format compatible with the OAK-D hardware. Firstly, VPU requires network weights to be downgraded from the commonly-used 32-bit floating-point precision (FP32) to 16-bit precision (FP16), which reduces the complexity and thus greatly increases the speed of the model, with the trade-off of reducing accuracy. Another challenge was that the batch size dimension of the input layer was originally saved as None, representing a dynamic input size. The VPU requires fixed sizes, so the rehabilitation model input layer was given a batch size of 1, since only one RGB image representation of BlazePose data will be fed to the network at a time. There was also a requirement for input image dimensions to be in planar/N-C-H-W order (Number of batches, Channels, Height of image, Width of image), as opposed to the interleaved/N-H-W-C order that was used in training. For this application, it involved reordering the input data from shape (1, 224, 224, 3) to (1, 3, 224, 224), which was done within the Blob Converter.

Depth-AI systems are configured as pipelines - communication workflows that connect separate processes running within the device and allow interfacing with a host computer. They consist of nodes, of which there are several types: Python scripts (Script), image manipulators (ImageManip), host input/output devices (XLinkIn/XLinkOut), and neural networks (NeuralNetwork). There are many pipeline examples provided within the Depth-AI GitHub repository, including an implementation of BlazePose [119] adapted from the OpenVINO model zoo.

4.3.2 Topology Normalisation

Below is a diagram of the BlazePose pose tracking skeleton topology. Due to the reliance of BlazePose on its preceding BlazeFace [96] algorithm, there are several face joints in the skeleton model which were not present in the topologies used for the MoVi and KIMORE datasets. In addition to this, BlazePose does

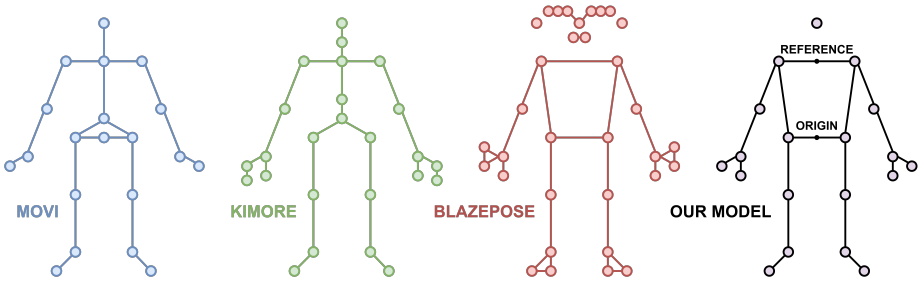


Figure 9: The skeletal topologies for pose estimation used by the MoVi and KIMORE datasets, in comparison to the BlazePose topology, and our normalised model. Our model consists of joints which are present for both KIMORE and BlazePose, with the origin set to the mid-hips and the skeleton scaled with respect to the distance between the new origin and a reference point - the mid-shoulders.

not track central torso joints, rather it uses the mid-hips and mid-shoulders as reference points depending on the visibility of the whole skeleton. Not only are the number and location of tracked joints different between topologies, but the order in which joints are listed also differs. This is crucial because the order in which joints are written to RGB determines their position within the neural network input image, and the network may have learned which pixels in the image are most indicative of a classification. To account for this, only the 19 common joints of the KIMORE and BlazePose topologies were kept, and the order was set to the BlazePose ordering system. The KIMORE toplogy, based on the Kinect V2, counts down each body part at a time (e.g. all the joints in the left arm, followed by the right arm) whereas the BlazePose topology counts each set of similar joints at a time (e.g. the left and right shoulder, followed by the left and right elbow).

Another factor to consider was the coordinates themselves; although both the KIMORE and BlazePose datasets measured 3D positions in metres, MoVi used millimetres, and all three frameworks had different axes, with respect to the ordering of x, y and z, the direction of positive and negative, and the location of the origin. Another factor to consider was that KIMORE data was collected with subjects stood at a fixed distance from the Kinect V2 sensor, which was also set up at a fixed height and angle. Therefore, the pose estimation data was normalised for each topology, and the networks were retrained with the new RGB images. Firstly, the coordinates of the mid-hips were set to the origin, with each subject scaled according to the distance between the mid-hips and mid-shoulders. In addition, all axes were set such that, from the view-point of the camera, the x-axis was increasing to the right, the y-axis was increasing upwards, and the z-axis was increasing towards the camera.

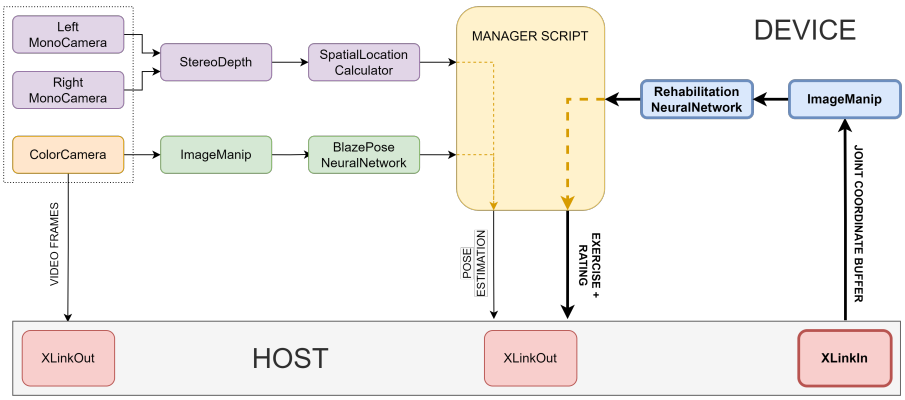


Figure 10: The complete AI camera pipeline, where the DepthAI BlazePose pipeline [119] has been merged with the telerehabilitation performance feedback system, highlighted in bold on the right side of the diagram. Arrows in and out of the XLink nodes represent the data queues between the host laptop and OAK-D device.

4.3.3 Integration with the BlazePose Pipeline

A diagram of the working telerehabilitation pipeline is shown in Figure 10. This diagram is an abstraction of the full pipeline graph (see Appendix), in which sub-systems have been organised by colour for clarity. The aim for integration of the rehabilitation network with the original Depth-AI BlazePose pipeline [119] was to take a 90-frame buffer of full-body joint position estimates, normalise the data, and reshape it into an RGB image for neural inference of the exercise and rating predictions, which are displayed on the host device.

Firstly, the purple sub-system contains depth estimation nodes. Captures from the left and right MonoCameras are passed to the StereoDepth node, which calculates depth with respect to frame disparity (the distance between corresponding points in the left and right frames), the focal length of the camera, and the physical separation distance of the MonoCameras within the OAK-D [120]. The StereoDepth node outputs depth maps, which are then used in the SpatialLocationCalculator node to estimate the spatial coordinates of the region-of-interest: for this application, it is the location of the mid-hips of the subject.

Although depth information from the OAK-D is utilised in the pipeline, the SpatialLocationCalculator node does not infer the z-positions of all 33 of the BlazePose key-points. The creators of the pipeline provide two reasons for this. Firstly, if key-point estimates from BlazePose are inaccurate, the coordinates provided for each joint may not align with their position on the stereo depth map. Secondly, z-positions can only be calculated if there are no occlusions. Therefore, the SpatialLocationCalculator is only used to calculate the depth of the mid-hips, which acts as the origin in the global coordinate

frame. This information is then merged with the inferred 3D landmarks from the pose estimation.

The ColorCamera, shown in orange, has two outputs: it sends captured video frames to the host computer, as well as RGB frames to the ImageManip node of the BlazePose sub-system, shown in green. ImageManip (Image Manipulation) nodes are used for cropping and resizing image frames. Since BlazePose is a top-down pose estimation algorithm, it requires a cropped image containing a single person to be passed to the network, before it regresses joint key-points, and therefore an ImageManip node is used to do this. The bounding box of the person is only provided if BlazePose detects a person in the frame - if this is the case, the coordinates of this box are passed to the ImageManip node as its input configuration (see Appendix). The BlazePose NeuralNetwork then regresses the locations of key-points within this image.

The Manager Script node, shown in yellow, contains a Python script which runs on-device and manages the control flow of data through the pipeline. It is responsible for setting input configurations for nodes, as well as processing NeuralNetwork outputs. The Manager combines the z-position of the mid-hips with the 3D coordinates from BlazePose, sending the resulting 'world landmark' coordinates to the host computer, alongside other pose information such as the confidence of each prediction. It also processes the outputs of our custom Rehabilitation NeuralNetwork.

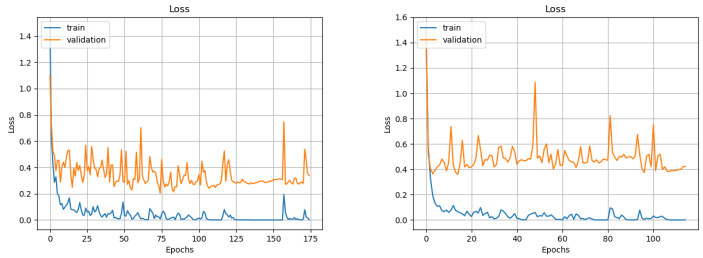
The XLinkOut and XLinkIn nodes, shown in red, control the device's outputs and inputs with respect to the host computer. The labelled arrows to and from the XLink nodes represent data queues. The video frame and pose queues are 'blocking' queues: other processes on the device are halted until data is provided via these XLinkOut nodes. In contrast, the queues for our application, containing rehabilitation network outputs, and joint coordinate inputs, are 'non-blocking', as data is only sent every 90 frames.

The telerehabilitation extension of the original pipeline has been shown in bold. An XLinkIn node transmits the 90-frame buffer of RGB-scaled, normalised joint coordinate estimates, of dimensions (3, 90, 19), to an ImageManip node, which resizes it to the ResNet-50 input shape of (3, 224, 224) using bi-linear interpolation, and passes it as input to the Rehabilitation NeuralNetwork. The network outputs two normalised vectors of shapes (1,5) and (1,3), representing the exercise and rating prediction probabilities. The original XLinkOut node from the Manager Script was modified to transmit the two vectors to the host.

For our application, the host computer was a commercial laptop operating with an Intel i7 core. In the telerehabilitaton pipeline, the host is tasked with checking that all 33 of the BlazePose key-points are provided for each frame, and that all joints have an acceptably high prediction confidence. This is equivalent to ensuring that the full body of the subject is in-frame. If this is the case, the host normalises the skeleton, using our model topology shown in Figure 9, and places the coordinates in a buffer, which collects data for 90 frames - this is the same number of frames used to train the network on KIMORE sequences. If tracking is lost at any point, the buffer is reset and the process is repeated. Once the buffer is full, The 3D coordinates are normalised to an RGB range of

[0, 255], and the resulting image is sent to the ImageManip node via XLinkIn.

The host laptop screen shows live video captures, with overlayed BlazePose skeletal tracking, to the subject in real-time, as well as most recent exercise and rating predictions. It also informs the subject if they are not in frame, or if certain joints are not visible to the sensor. Due to the 90-frame buffer, and the system's operating frame-rate of 20fps, there is a lag of around 4-5 seconds between feedback messages to the subject. However, this is a sufficiently short period for real-time applications: if the subject receives a low rating, they can adjust their exercise performance, and receive updated feedback on their motion within seconds, to determine if they have made an improvement. Most data processing tasks were implemented on the host, rather than the OAK-D, because the Depth-AI Python packages, as well as NumPy and other libraries required for data manipulation, are not available on the device's hardware. Additionally, the VPU processor is tailored for neural network computations, and has little available memory for CPU operations. Since the tracking data was already being sent to the host, and data queues are transmitted at high speed via USB3 cable, the additional task of buffering the RGB images on-device was deemed to be negligible, and has had no noticeable effect on the run-time of the system.



(a) Non-normalised topology (b) Normalised topology

Figure 11: A comparison of the network loss graphs during training, in which early stopping was implemented, for MoVi sequences with and without normalisation applied.

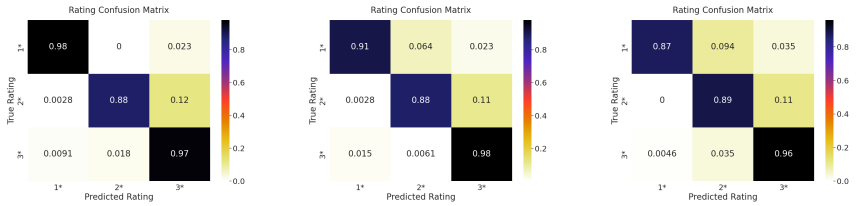
5 Results and Discussion

5.1 The Rehabilitation Neural Network

The ResNet-50 network was trained both with non-normalised MoVi sequences, containing data for 20 tracked joints, before the normalised model was implemented, in which the MoVi skeleton had its torso joints removed, reducing the key-point count to 17, and had its scale normalised, according to the topology diagram in Figure 9. With both sets of data, and the same configuration, the network achieved over 95% classification accuracy. This was considered to be sufficiently high for the use-case of transfer learning. While the loss was also low, there was a discrepancy between the models, with the non-normalised model loss reaching 0.2, while the normalised model loss was 0.4, as shown in Figure 11.

Normalisation involved the scaling of all subjects such that the distance from mid-hips to mid-shoulders was 1. This removed some information contained within their skeleton that the network may have exploited for inference, such as their height. In addition, the normalised MoVi skeleton does not contain the pelvis, mid-hips or spine joints, as these are not provided with BlazePose. It is likely that these joints were useful for activity inference, and it may have been more challenging for the network to classify activities without them. However, the difference is negligible in comparison to the detrimental impact that training on incompatible topologies would have. Normalisation made the problem of classification slightly more challenging, with the trade-off of better model generalisation for later training stages.

We trained the rehabilitation network with three sets ResNet-50 configurations: our method with MoVi pre-training, ResNet-50 with imported ImageNet weights [116], and ResNet-50 with randomly-initialised weights. The reason for doing this was to test our hypothesis that pre-training on high-resolution activity sequences would improve the classification accuracy of rehabilitation data captured with a lower-resolution sensor.



(a) MoVi pre-trained weights (b) ImageNet weights (c) Random weights

Figure 12: Confusion matrices representing rating classification accuracy from the KIMORE-trained rehabilitation network, using three different ResNet-50 weight initialisations, on a test split containing 20% of the samples.

The KIMORE dataset was split in such a way that 20% of the sequences were kept for testing, while the remaining 80% were used for training and validation. In addition, the test set exclusively contained subject data that the network had not been trained on, and the rating distribution of sequences was proportional to the overall distribution of each rating class in the dataset.

The confusion matrices for rating classification are shown in Figure 12. The results were comparable for 2* and 3* rating classification, with all networks achieving a high accuracy with these scores, but a noticeable difference for the 1* rating class, representing the poorest performances contained in the dataset.

It is important to note that accuracy alone is not a reliable method for assessing the performance of an imbalanced multi-class classification network. Two other metrics for assessing classification are precision - the proportion of class predictions that are correct - and recall - the proportion of true values that are correctly classified. For this implementation, recall is of greater priority, as false positives (incorrectly labelling a good performance as bad) are less dangerous in practice than false negatives (incorrectly labelling a bad performance as good).

The recall scores for rating classification, generated by scikit-learn [], are shown in Table 1. As with the results displayed in Figure 12, scores are similar for 2* and 3* classification, but pre-training ResNet-50 with the MoVi activity recognition dataset demonstrated a clear advantage for the correct identification of poor performances.

	Classification Recall Score			
ResNet-50 Initialisation	1*	2*	3*	Weighted Average
MoVi Pre-trained weights	**0.977**	0.875	0.972	**0.943**
ImageNet weights	0.912	0.883	**0.977**	0.940
Random weights	0.871	**0.895**	0.961	0.923

Table 1: The rating classification recall scores for the rehabilitation network for each of the three ResNet-50 weight initialisations.

However, there are some limitations to our approach for rating classification. KIMORE dataset scores are provided as an overall assessment of the subject's exercise sequence, which comprises of five motion repetitions. In our application, only 3 seconds from each sequence are fed to the network at a time, with the sample being assigned the same rating as the entire sequence. It is likely than in practice, the real score of the subject would fluctuate between repetitions. This was a key reason why we decided to abstract the score into a rating classification system: it was theorised that good performances would likely be good for the whole sequence, while poor performances, most often from real rehabilitation patients, were poor for the whole sequence. However, this may not have always been the case, and so the network may have learned in some cases to identify slight erroneous movements as good performances, if they were, for example, performed by a healthy subject within a high-scoring sequence.

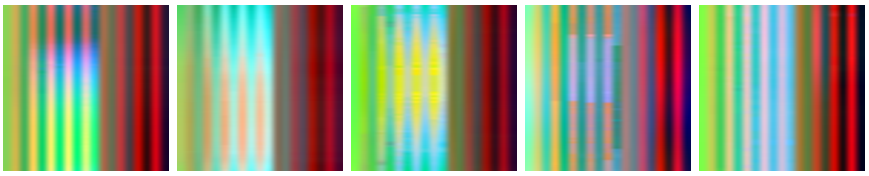
5.2 AI Camera Integration

The 'Lite' version of the DepthAI BlazePose implementation has an advertised frame-rate of 22fps [119]. In comparison, our pipeline, with the telerehabilitation network, operates at 20fps. This performance is superior to the OAK-D rehabilitation pipeline in [24], which advertised 15fps with the OpenPose pose estimation algorithm, with an Intel Core i7 host laptop - the same as used in our application. In addition, OpenPose only provides coordinates for 19 tracked joints in 2D, in comparison to BlazePose's 33 tracked key-points in 3D. While there is a slight accuracy trade-off between OpenPose and BlazePose Lite [34], the difference in run-time makes the latter more suitable for dynamic motion applications.

The AI camera pipeline was tested with a proof-of-concept approach, since a thorough review of its accuracy would require a clinician to provide feedback on the subject's performances as a ground-truth for comparison with the rehabilitation network's rating outputs.

We assessed the ability of the AI camera pipeline to differentiate between exercises and provide realistic performance ratings by having a subject complete repetitions of the KIMORE rehabilitation exercises, shown in Figure 3, in front of the sensor. The OAK-D was mounted at a height of 1m, with its line-of-sight at a 90° angle to the floor. The subject was stood 3m from the sensor, which was the distance used for KIMORE dataset capture. This separation allowed for the whole body to be within the sensor's field-of-view for all exercise sequences, including those with arm extensions above the head.

Implementation of the data recording process was done using a finite state machine, which ran on the host computer. A 90-frame 'WAIT' state was introduced between exercises, for which the system would not collect pose information to feed into the rehabilitation network. This state was also entered whenever BlazePose was unable to locate all of the subject's joints. The 'WAIT' state was communicated via the laptop interface, alongside an indication of whether their whole body was within frame. This allowed the subject to adjust their position in case of occlusions or lost tracking. Additionally, a 'RECORD' state was used to capture 900 frames, corresponding to 10 RGB ResNet-50



(a) Exercise 1 (b) Exercise 2 (c) Exercise 3 (d) Exercise 4 (e) Exercise 5

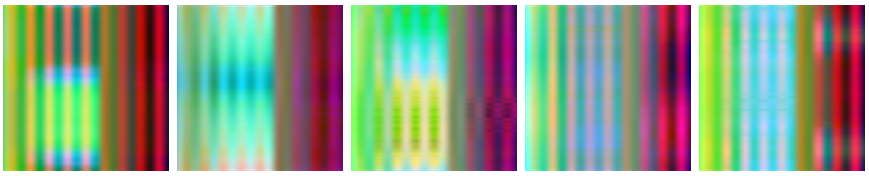
Figure 13: Normalised RGB images representing 90-frame samples of each exercise performed by an expert subject (E_ID2) in the KIMORE dataset.

inputs and therefore 10 sets of predictions, for each exercise. During this state, both the exercise and rating prediction were displayed on the laptop interface, allowing the subject to not only ensure that the system was tracking their motion, but to also allow them to adjust their performance according to feedback from the system, similarly to how a commercial implementation would work. Finally, once all exercises were recorded and analysed, the program reached a 'STOP' state, in which the device was switched off and all information collected throughout the experiment was written to file. The system also saved each RGB input, to enable comparison with images generated from the KIMORE dataset. As well as the prediction, representing the highest-probability element in the results vector, the raw neural network outputs were also saved, to allow for the comparison of prediction probabilities for each class.

The KIMORE dataset paper provides a thorough review of each exercise, including the starting position, posture, and other control factors, such as which joints to keep still and which plane to move in. Two experiments were conducted: firstly, the subject was asked to perform the motions while following the instructions for each exercise to the best of their ability, and secondly, they were asked to perform the motions incorrectly, such as with poor posture, slack arms, and movement of the hands and feet.

Some example images of each exercise generated from the KIMORE dataset are shown in Figure 13. In comparison, images generated by our pipeline's ImageManip node, representing the test subject performing each exercise in front of the OAK-D, are shown in Figure 14. Despite being captured with different sensor technologies, running different pose estimation algorithms, normalisation has ensured that sequences are as similar as possible. This indicates that the rehabilitation network is compatible for use with the OAK-D, without the need to re-train the network on BlazePose sequences. The only noticeable difference between these sets of captures is in exercise 4 - pelvis rotations in the transverse plane - where a ripple in the middle columns of the image in Figure 13(d), indicating a movement in the torso and hip joints, is not present in Figure 14(d).

Example captures from the 'RECORD' state are shown in Figure 15. Figure 15(a) shows the correct classification of Exercise 1: arm lift, with a rating score of 2*. Although this sequence was recorded while the subject was attempting



(a) Exercise 1 (b) Exercise 2 (c) Exercise 3 (d) Exercise 4 (e) Exercise 5

Figure 14: Normalised RGB images representing 90-frame samples of each exercise performed as well as possible by our test subject, in front of the AI camera.

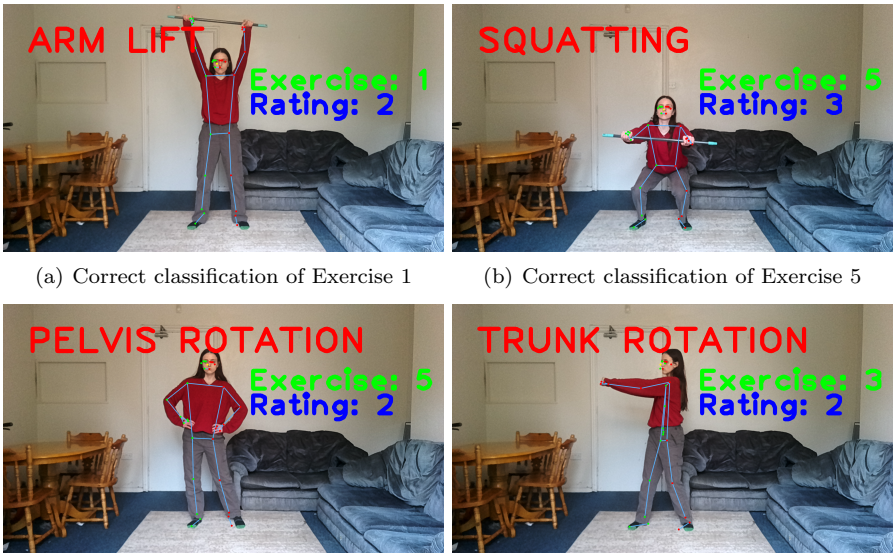
the exercises as well as possible, it can be seen in the frame capture that their left arm was slightly skewed from their right arm. As this exercise specifies that both arms must be completely extended, this may be the reason behind this sequence's rating classification. Additionally, Figure 15(b) shows the correct classification of Exercise 5: squatting, which was awarded a score of 3*. In comparison, Figure 15(c) shows an example of an incorrect classification of Exercise 4: pelvis rotations. Reasons for the system's difficulty with this exercise are provided in section 5.2.1. Finally, Figure 15(d) demonstrates a failure of the BlazePose tracking algorithm, which incorrectly predicts the location of the right arm, due to its occlusion by the left arm. Despite this occlusion occurring several times during this sequence, the exercise classification was unaffected. One possible reason for this is that the KIMORE sequences also contained instances of lost arm tracking, and so the network learned to be robust to these errors for the trunk rotation sequence. Another reason may be that this motion is sufficiently different enough from the other exercises to prevent misclassification.

When the subject was instructed to perform each exercise as accurately as possible, the expected results were high rating scores for each exercise, and for all exercises to be classified correctly. However, it is important to note that since there was no professional present to supervise the motions, it is likely that the subject made some mistakes during the experiment, as suggested in Figure 15(a), and therefore the ground-truth ratings may not be 3* for all sequences. Conversely, the performances with intentional errors were expected to achieve low scores, but a lack of ground-truth prevents us from being able to determine if higher rating classifications are incorrect.

Therefore, while the exercise classifications can be compared to a ground-truth for validation of our system, the rating scores are provided as an indication of the network's prediction patterns, rather than a test of the system's accuracy for performance feedback.

5.2.1 Exercise Classification

The confusion matrices for both the well-performed and poorly-performed exercise sequences are shown in Figure 16. For the sequences performed as well as



(a) Correct classification of Exercise 1

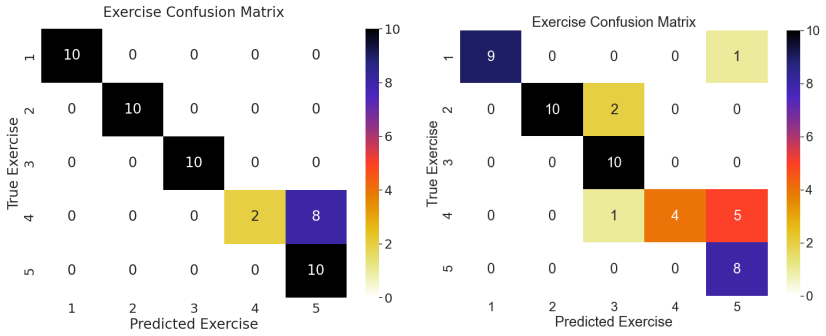
(b) Correct classification of Exercise 5

(c) Incorrect classification of Exercise 4

(d) Incorrect BlazePose tracking of the right arm for Exercise 5

Figure 15: Example captures from the laptop interface of the working rehabilitation pipeline during experimental data collection. The exercise instruction is shown in red, with the neural network predictions displayed in green and blue.

possible, the network achieved perfect classification accuracy for all but one of the exercises. In comparison, the network was more likely to mis-classify exercises that were poorly performed. Well-performed sequences comprise the majority of the training data, so it was expected that the network would find them easier to identify. In addition, there is only one way to perform an exercise correctly, due to the strict instructions on posture, speed of motion, and joint angles to maintain, whereas there are many ways in which a motion can be performed badly, with examples including poor posture, movement of the arms and legs, torso tilt, slack limbs, and variable movement speeds. The way in which the healthy test subject performed exercises incorrectly is likely to differ from that of real patients, who comprise the majority of the 1* sequences used to train the network.



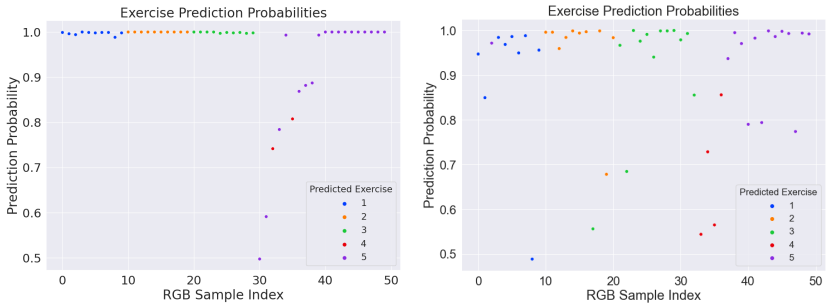
(a) Good performance

(b) Poor performance

Figure 16: Confusion matrices showing the classification accuracy of the neural network for test subject sequences of exercises performed as well as possible, in comparison to exercises performed with intentional errors.

Surprisingly, for both experiments, the network failed to identify instances of Exercise 4: pelvis rotations in the transverse plane, and was more likely to classify these sequences as Exercise 5: squatting. As a further investigation of this, Figure 17 shows a scatter plot of the results of exercise classification, with respect to the probability of each classification assignment, corresponding to the degree of confidence that the network had for each prediction. The graphs indicate that for the well-performed sequences, the prediction probabilities of correctly-identified exercises was high, implying that the network was confident in these classifications. However, there was also a relatively high confidence in its incorrect classifications of Exercise 4. While the overall confidence in exercise classification was lower for poor performances, and there was an occasional mis-classification, the network was more likely to classify Exercise 4 correctly for this experiment, although its confidence in these predictions was low.

Our initial hypothesis for the poor classification results for Exercise 4 was that the pipeline's z-coordinate tracking may be unsuitable for capturing the hip joints' small trajectories of motion in the transverse (X-Z) plane. As discussed in



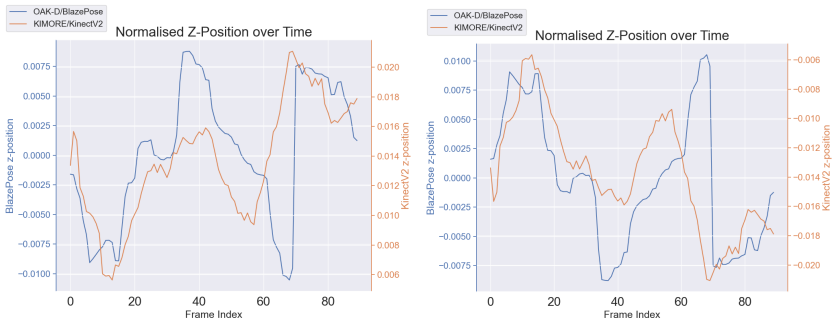
(a) Good exercise performance (b) Poor exercise performance

Figure 17: A scatter plot of the prediction probabilities of each exercise classification generated by the rehabilitation network, for test subject sequences of exercises performed as well as possible, in comparison to exercises performed with intentional errors.

Section 4.3.3, BlazePose is used to estimate the 3D positions of all tracked key-points, while the OAK-D's depth sensing technology is only used to calculate the z-coordinate of the mid-hips. In comparison, the KIMORE dataset was captured using the Kinect V2, which uses time-of-flight to regress the z-coordinates of every tracked joint, rather than a single point, when predicting 3D pose. In addition, the OAK-D's depth accuracy decreases exponentially as the distance between the sensor and its region-of-interest increases [120].

To determine if the pipeline's z-coordinate tracking was causing these results, the left and right hip joints of the test subject, captured during the well-performed sequence, were plotted against the same joints from a 3* KIMORE dataset sample. The trajectories are shown in Figure 18. Although the plots represent the coordinates of different subjects, captured with different pose estimation algorithms running on different sensors, the pattern of movement appears to correlate quite closely, showing the oscillations of the hips as the pelvis rotates. This suggests that despite our initial hypothesis, the pipeline may be sufficiently capturing the small trajectories of motion in the hips, even though the depth estimation method is inferior to that of the Kinect.

Therefore, we considered a few other reasons why the network may have confused these two exercises. Firstly, Exercises 4 and 5 focus movement on the lower body, whereas the first three exercises are concentrated on the arms and upper torso. This may be a reason why pelvis rotations, which require the placement of hands on hips, were falsely classified as squats. In addition, it is possible that the network learned to recognise pelvis rotations from a different set of joints. This hypothesis is backed up by the RGB images in Figures 13 and 14: Exercise 4 shows a visible motion ripple in 13(d) that is not present in 14(d), for the middle columns of the image, corresponding to the indexes of arm joints. Another factor to consider is that for some squatting sequences in the KIMORE dataset, the subject placed their hands on their hips, instead of



(a) Left hip (b) Right hip

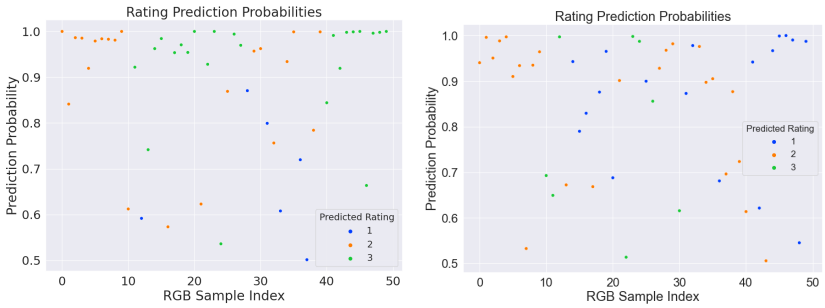
Figure 18: A comparison of the normalised z-positions of the left and right hip joints for a pelvis rotation sequence, captured by the Kinect V2 (orange) and BlazePose (blue) pose estimation algorithms.

holding a bar perpendicular to their torso. As the network was trained on RGB images, rather than kinematic constraints such as joint angles and velocities of motion, it may be the case that ResNet-50 found it easier to classify Exercise 4 from pixels corresponding to the trajectory of the arms, which have a larger range of motion than the hips. If this is the case, it is likely that the arm positions were also more likely to influence the rating classification, despite the KIMORE paper stating that the motion's correctness was determined by the pelvis trajectory. Further work on this topic may seek to identify which joints the network pays the greatest attention to when classifying sequences.

5.2.2 Rating Classification

Figure 19 shows the rating classifications generated for each 90-frame RGB sample, with respect to the probability, or confidence, of each score. As can be seen, some exercises were classified as being majority 3* for their duration, and others never reached a 3* rating. As previously mentioned, the accuracy of these results can't be properly ascertained without professional supervision of the subject's motions, and it may be the case that the subject's performances of the motions would have been awarded a low clinical score in practice. However, this plot shows that most samples were given a 2* or 3* rating, which is realistic given that the subject, who has no known injuries or conditions, was performing the sequences as well as possible during data collection. The plot also shows that some exercises were more likely to be awarded higher ratings than others - exercise 1, for example, was confidently awarded 2* throughout, while the network awarded 3* with high probability for most of exercise 5. It is difficult to determine the reason for this difference without a clinician to provide a ground-truth quality assessment.

In the set of results for which the subject performed each motion incorrectly on purpose, the neural network was more likely to mis-classify the exercise. The



(a) Good exercise performance

(b) Poor exercise performance

Figure 19: A scatter plot representing the prediction probability of each rating classification generated by the rehabilitation network, for test subject sequences of exercises performed as well as possible, in comparison to exercises performed with intentional errors.

network also provided fluctuating rating scores, with only a few exercises being awarded consistently low ratings. This indicates that in a real application, a feedback connection might be useful when the rating is low, such that the network is less likely to immediately award a high rating for a similar motion. Another way to implement this would be to only award ratings of 3* when the network's confidence in both the exercise and rating are high. If the probabilities of two rating classes are similar, it would be beneficial to choose the lower rating, as an incorrect positive feedback would hinder recovery in real applications.

Another factor to consider is that in the KIMORE dataset, the subjects were all instructed to perform exercises as well as possible. As a result of this, most of the sequences which were awarded low scores were performed by real rehabilitation patients, who suffer from musculoskeletal disorders which the exercises are designed to correct. In contrast, the subject used to test our system has no known health conditions, so to simulate poor performances, the exercises had to be completed with intentional errors. Therefore, it is likely that the neural network struggled to classify these sequences because it has been trained to identify poor performances from real patient data, rather than of healthy subjects with intentional bad posture. To determine the system's applicability for a rehabilitation setting, it should be tested in a medical environment with real patients, with the rating classifications assessed with respect to ground-truth clinical scores.

The KIMORE paper specifies that for many exercises, an incorrect motion may involve tilting of the trunk or pelvis, or a lack of proper alignment in the arms. As previously described, the OAK-D's depth technology is not as accurate of the Kinect's time-of-flight sensor. This means that the rehabilitation network, trained on sequences captured with the Kinect, may have learned to exploit z-coordinate information as a key identifier of poor performances. Small motions in the transverse plane are more likely to be detected with Kinect than

the OAK-D, and so for some exercises, where the Primary Outcome score is high, indicating a correct overall motion, but the Control Factor score is low, indicating a poor posture, the OAK-D might not be picking up on the depth information necessary to classify the sequence with a low rating. This is another indication that an AI camera with a superior depth sensor, such as the OAK-D Pro, might be required. However, the strong classification results for four out of five of the exercises indicate that the OAK-D AI camera is a feasible platform for many rehabilitation applications which primarily involve sagittal and frontal movement, such as KIMORE exercises 1, 2 and 3.

6 Conclusion

To evaluate the outcome of this thesis, each of the objectives stated in section 2.2 are considered in turn.

Firstly, pre-training an image classification architecture on human activity sequences was successful, with the network reporting a sufficiently high accuracy and low loss, despite the loss of skeletal information due to normalisation. The benefit of pre-training was ascertained by its high recall score for the classification of sequences awarded a 1* rating, as shown in Table 1.

Secondly, the re-trained network, based on KIMORE rehabilitation sequences, displayed strong results for exercise and rating classification accuracy on test data. In addition, the use of topology normalisation allowed for the network to successfully classify BlazePose exercises captured by the OAK-D, despite a lack of these sequences in the training data. While the accuracy for rating classification with BlazePose sequences was not able to be determined, the results suggested that the network was able to roughly differentiate between good and bad performances. However, the poor identification of pelvis rotation sequences indicates a flaw in the system, which may be due to the network's focus on unexpected joints for exercise classification.

Finally, the integration of the rehabilitation network with an existing pose estimation pipeline led to a system capable of achieving the thesis aim: to create a platform for autonomous, real-time rehabilitation, capable of providing realistic feedback. Furthermore, the addition of our network to the BlazePose pipeline resulted in a decrease in frame-rate of just 2fps. A 20fps pose estimation algorithm for real-time feedback, running on a commercial laptop, is more than sufficient for capturing dynamic movements.

In conclusion, we have demonstrated the feasibility of AI cameras as a platform for telerehabilitation, especially as a low-cost alternative to commercial RGB-D sensors. In practice, this sensor could be paired with a Raspberry Pi and an LCD screen as a "plug-and-play" kit, requiring minimal set-up time for patients, and thus improving usability.

6.1 Future Work

The next step would be to properly test the system in a clinical environment, to determine the accuracy of rating classification. This should be done with many subjects, each performing several repetitions of each exercise, and with multiple clinicians present who are filling out the questionnaire provided in the KIMORE paper and providing a score for each set of sequences. The ratings would then be provided after the sequence is complete, using the same abstraction as the neural network was trained on.

As well as improving the current implementation, there are several ways in which it could be extended to provide more detailed feedback to the subject. The KIMORE dataset provides scores as a combination of control factors and primary outcomes, which were not considered for this application. However, if the overall score is low, these individual sub-scores are a good indication of what

the issue could be. For example, the Primary Outcome score being low would suggest that the exercise motion itself is being performed incorrectly, whereas a low Control Factor score would indicate that the issue lies with the subject's posture and stance. Each exercise in the dataset has a different set of primary outcomes and control factors, specific to the parts of the body targeted by the exercise.

An even more specific form of feedback would be to tell the subject which joints are the most responsible for their poor exercise performance. The structure of each RGB input image to the rehabilitation network, in which columns represent joints and are ordered in a top-down approach, could be exploited in such a way that the network can output which part of the image, and therefore which body parts, are most responsible for a given score. This would be particularly useful for the pelvis rotation exercise. A simple way of implementing this is with a Grad-CAM class activation visualisation [121]. A heat-map of the RGB image, indicating the pixels which most affected classification, would be analysed. For example, if the arms are slack, the pixels covering the arm joints would show up with Grad-CAM, and the user could be notified on the interface not only of their low rating, but that their arms are causing the issue. This feedback would be a lightweight addition to the current architecture, unlikely to hinder frame-rate, and would greatly improve the user experience, as the subject would know what to do to fix their posture and raise their score. s

References

[1] A. Cieza, K. Causey, K. Kamenov, S. W. Hanson, S. Chatterji, and T. Vos, "Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10267, pp. 2006–2017, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673620323400

[2] S. Negrini, "The World Health Organization "Rehabilitation 2030: a call for action"." 4 2017.

[3] S. Machlin, J. Chevan, W. Yu, and M. Zodet, "Determinants of Utilization and Expenditures for Episodes of Ambulatory Physical Therapy Among Adults," *Physical therapy*, vol. 91, pp. 1018–1029, 1 2011.

[4] R. Komatireddy, "Quality and Quantity of Rehabilitation Exercises Delivered By A 3-D Motion Controlled Camera: A Pilot Study," *International Journal of Physical Medicine & Rehabilitation*, vol. 02, 1 2014.

[5] E. Sluijs, G. Kok, and J. Zee, "Correlates of Exercise Compliance in Physical Therapy," *Physical therapy*, vol. 73, pp. 771–82, 1 1993.

[6] K. Jack, S. Mclean, J. Moffett, and E. Gardiner, "Barriers to treatment adherence in physiotherapy outpatient clinics: A systematic review," *Manual therapy*, vol. 15, pp. 220–228, 1 2010.

[7] W. H. Organization, "Adherence to long-term therapies : evidence for action," p. 196 p., 2003.

[8] B. Debnath, M. O'Brien, M. Yamaguchi, and A. Behera, "A review of computer vision-based approaches for physical rehabilitation and assessment," *Multimedia Systems*, vol. 28, pp. 209 – 239, 2021.

[9] H. Mousavi Hondori and M. Khademi, "A Review on Technical and Clinical Impact of Microsoft Kinect on Physical Therapy and Rehabilitation," *Journal of Medical Engineering*, vol. 2014, pp. 1–16, 12 2014.

[10] T. Richmond, C. Peterson, J. Cason, M. Billings, E. Terrell, A. Lee, M. Towey, B. Parmanto, A. Saptano, E. Cohn, and D. Brennan, "American Telemedicine Association's Principles for Delivering Telerehabilitation Services," *International Journal of Telerehabilitation*, vol. 9, pp. 63–68, 1 2017.

[11] H. Mousavi Hondori, "A Review on Technical and Clinical Impact of Microsoft Kinect on Physical Therapy and Rehabilitation," *Journal of Medical Engineering*, vol. 2014, 2 2014.

[12] Microsoft, "Azure Kinect DK Documentation." [Online]. Available: https://docs.microsoft.com/en-us/azure/kinect-dk/

[13] OpenCV, "OpenCV AI Kit with Depth." [Online]. Available: https://docs.luxonis.com/projects/hardware/en/latest/pages/BW1098OAK.html

[14] D. Lu, "AI-powered smartphone cameras are changing the way we see reality," *New Scientist*, 3 2019. [Online]. Available: https://www.newscientist.com/article/mg24132214-500-ai-powered-smartphone-cameras-are-changing-the-way-we-see-reality/

[15] "Lights, camera, action: The rise of AI cameras in transport," *SMMT*, 3 2023. [Online]. Available: https://www.smmt.co.uk/2023/03/lights-camera-action-the-rise-of-ai-cameras-in-transport/

[16] R. Clancy, "How AI cameras can improve commutes," *Electronics360*, 2022. [Online]. Available: https://electronics360.globalspec.com/article/18946/how-ai-cameras-can-improve-commutes

[17] B. Dickson, "Is camera-only the future of self-driving cars?" *ADAS & Autonomous Vehicle International*, 5 2022. [Online]. Available: https://www.autonomousvehicleinternational.com/features/is-camera-only-the-future-of-self-driving-cars.html

[18] D. Whittaker, "Why AI CCTV is the future of security and surveillance in public spaces," *Security Magazine RSS*, 12 2021. [Online]. Available: https://www.securitymagazine.com/articles/96719-why-ai-cctv-is-the-future-of-security-and-surveillance-in-public-spaces

[19] A. Bajaj, "The power of Artificial Intelligence in Drones," *Analytics Vidhya*, 7 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/07/the-power-of-artificial-intelligence-in-drones/

[20] Z. Ma, R. Rayhana, Z. Liu, G. G. Xiao, Y. Ruan, and J. S. Sangha, "Industrial internet of things (IOT) and 3D reconstruction empowered Smart Agriculture System," *2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 2022.

[21] S. Gunturu, A. Munir, H. Ullah, S. Welch, and D. Flippo, "A Spatial AI-Based Agricultural Robotic Platform for Wheat Detection and Collision Avoidance," *AI*, vol. 3, no. 3, pp. 719–738, 2022. [Online]. Available: https://www.mdpi.com/2673-2688/3/3/42

[22] T. Orsag Luka
}and Stipancic, K. Leon, and P. Karlo, "Human Intention Recognition for Safe Robot Action Planning Using Head Pose," in *HCI International 2022 - Late Breaking Papers. Multimodality in Advanced Interaction Environments*, S. Kurosu Masaaki

}and Yamamoto, M. Hirohiko, S. D. D., F. C. M., S. N. A., and K. Shin'ichi, Eds. Cham: Springer Nature Switzerland, 2022, pp. 313–327.

[23] A. Raza, M. H. Yousaf, and S. A. Velastin, "Human Fall Detection using YOLO: A Real-Time and AI-on-the-Edge Perspective," in *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, 2022, pp. 1–6.

[24] D. Perazzo, N. Soares, V. Lyra, G. Lima, A. Da Gama, J. Teixeira, and V. Teichrieb, "OAK-D as a Platform for Human Movement Analysis: A Case Study," 1 2021, pp. 167–171.

[25] "Intel® Movidius™ Myriad™ X Vision Processing Unit (VPU)." [Online]. Available: https://www.intel.co.uk/content/www/uk/en/products/details/processors/movidius-vpu/movidius-myriad-x.html

[26] J. Sandino, J. Galvez-Serna, N. Mandel, F. Vanegas, and F. Gonzalez, "Autonomous mapping of desiccation cracks via a probabilistic-based motion planner onboard uavs," *2022 IEEE Aerospace Conference (AERO)*, 2022.

[27] N. Mandel, J. Sandino, J. Galvez-Serna, F. Vanegas, M. Milford, and F. Gonzalez, "Resolution-Adaptive Quadtrees for semantic segmentation mapping in UAV applications," *2022 IEEE Aerospace Conference (AERO)*, 2022.

[28] L. O. Rojas-Perez and J. Martinez-Carranza, "DeepPilot4Pose: A Fast Pose Localisation for MAV Indoor Flight Using the OAK-D Camera," *J. Real-Time Image Process.*, vol. 20, no. 1, 2 2023. [Online]. Available: https://doi.org/10.1007/s11554-023-01259-x

[29] T. Simpson, "Real-Time Drone Surveillance System for Violent Crowd Behavior Unmanned Aircraft System (UAS) – Human Autonomy Teaming (HAT)," *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pp. 1–9, 2021.

[30] L. O. Rojas-Perez and J. Martinez-Carranza, "Towards autonomous drone racing without GPU using an oak-D smart camera," *Sensors*, vol. 21, no. 22, p. 7436, 2021.

[31] T. Wakayama, G. A. Garcia Ricardez, L. El Hafi, and J. Takamatsu, "6d-pose estimation for manipulation in retail robotics using the inference-embedded oak-D camera," *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022.

[32] J. A. Sánchez-Rojas, J. A. Arias-Aguilar, H. Takemura, and A. E. Petrilli-Barceló, "Staircase Detection, Characterization and Approach Pipeline for Search and Rescue Robots," *Applied Sciences*, vol. 11, no. 22, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/22/10736

[33] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity }Fields," *CoRR*, vol. abs/1812.08008, 2018. [Online]. Available: http://arxiv.org/abs/1812.08008

[34] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," *CoRR*, vol. abs/2006.10204, 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[35] E. der Kruk and M. Reijne, "Accuracy of human motion capture systems for sport applications; state-of-the-art review," *European Journal of Sport Science*, vol. 18, pp. 1–14, 4 2018.

[36] I. Stancic, T. G. Supuk, and A. Panjkota, "Design, development and evaluation of optical motion-tracking system based on active white light markers," *IET Science, Measurement & Technology*, vol. 7, no. 4, pp. 206–214, 4 2013. [Online]. Available: https://bris.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/design-development-evaluation-optical-motion/docview/1492949756/se-2

[37] Y. Song and I. Biro, "The Evolution of Marker-Based Motion Analysis and the Integration of Advanced Computational Methods: Application to Human Gait Biomechanics," in *2022 2nd International Conference on Bioinformatics and Intelligent Computing*, ser. BIC 2022. New York, NY, USA: Association for Computing Machinery, 2022, pp. 201–205. [Online]. Available: https://doi.org/10.1145/3523286.3524689

[38] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, and X. Savatier, "A Study of Vicon System Positioning Performance," *Sensors (Basel, Switzerland)*, vol. 17, 2017.

[39] M. Windolf, N. Goetzen, and M. M. Morlock, "Systematic accuracy and precision analysis of video motion capturing systems–exemplified on the Vicon-460 system." *Journal of biomechanics*, vol. 41 12, pp. 2776–80, 2008.

[40] A. Panjkota, I. Stančic, and T. Šupuk, "Outline of a Qualitative Analysis for the Human Motion in Case of Ergometer Rowing," in *Proceedings of the 9th WSEAS International Conference on Simulation, Modelling and Optimization*, ser. SMO'09. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2009, pp. 182–186.

[41] J. Lallemand, M. Szczot, and S. Ilic, "Human pose estimation in stereo images," in *Articulated Motion and Deformable Objects: 8th International Conference, AMDO 2014, Palma de Mallorca, Spain, July 16-18, 2014. Proceedings 8*, 2014, pp. 10–19.

[42] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose }Estimation from a Single Depth Image," *CoRR*, vol. abs/1908.09999, 2019. [Online]. Available: http://arxiv.org/abs/1908.09999

[43] T. Yu, J. Zhao, Z. Zheng, K. Guo, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "DoubleFusion: Real-Time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 10, pp. 2523–2539, 10 2020.

[44] J. Chen, H. Shi, Y. Ye, K. Yang, L. Sun, and K. Wang, "Efficient Human Pose Estimation via 3D Event Point Cloud," 2022.

[45] T.-Q. Wang, Y. You, K. Osawa, M. Shimodozono, and E. Tanaka, "A Remote Rehabilitation and Evaluation System Based on Azure Kinect," *Journal of Robotics and Mechatronics*, vol. 34, no. 6, pp. 1371–1382, 2022.

[46] A. K. Roy, Y. Soni, and S. Dubey, "Enhancing effectiveness of motor rehabilitation using kinect motion sensing technology," in *2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS)*, 2013, pp. 298–304.

[47] B. D. S. Gonçalves, O. Postolache, and J. M. D. Pereira, "Gait Rehabilitation in Virtual Reality Serious Game Interactive Scenarios," in *2022 International Conference and Exposition on Electrical And Power Engineering (EPE)*, 2022, pp. 672–676.

[48] Y. Xu, M. Tong, W.-K. Ming, Y. Lin, W. Mai, W. Huang, and Z. Chen, "A depth camera–based, task-specific virtual reality rehabilitation game for patients with stroke: Pilot usability study," *JMIR serious games*, vol. 9, no. 1, p. e20916, 2021.

[49] Y. Ling, L. Ter Meer, Z. Yumak, and R. Veltkamp, "Usability Test of Exercise Games Designed for Rehabilitation of Elderly Patients After Hip Replacement Surgery: Pilot Study," *JMIR Serious Games*, vol. 5, p. e19, 5 2017.

[50] C. Gu, W. Lin, X. He, L. Zhang, and M. Zhang, "IMU-based motion capture system for rehabilitation applications: A systematic review," *Biomimetic Intelligence and Robotics*, vol. 3, no. 2, p. 100097, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667379723000116

[51] D. Roetenberg, H. Luinge, and P. Slycke, "Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors," *Xsens Motion Technol. BV Tech. Rep.*, vol. 3, 4 2009.

[52] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popović, "Practical Motion Capture in Everyday Surroundings," *ACM Trans. Graph.*, vol. 26, no. 3, p. 35–es, 7 2007. [Online]. Available: https://doi.org/10.1145/1276377.1276421

[53] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs," *Comput. Graph. Forum*, vol. 36, no. 2, pp. 349–360, 5 2017. [Online]. Available: https://doi.org/10.1111/cgf.13131

[54] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model," *ACM Trans. Graph.*, vol. 34, no. 6, 11 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818013

[55] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time," *ACM Trans. Graph.*, vol. 37, no. 6, 12 2018. [Online]. Available: https://doi.org/10.1145/3272127.3275108

[56] X. Yi, Y. Zhou, and F. Xu, "TransPose: Real-Time 3D Human Translation and Pose Estimation with Six Inertial Sensors," *ACM Trans. Graph.*, vol. 40, no. 4, 7 2021. [Online]. Available: https://doi.org/10.1145/3450626.3459786

[57] Y. Tian, X. Meng, D. Tao, D. Liu, and C. Feng, "Upper limb motion tracking with the integration of IMU and Kinect," *Neurocomputing*, vol. 159, pp. 207–218, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231215001472

[58] Y.-C. Du, C.-B. Shih, S.-C. Fan, H.-T. Lin, and P.-J. Chen, "An IMU-Compensated Skeletal Tracking System Using Kinect for the Upper Limb," *Microsyst. Technol.*, vol. 24, no. 10, pp. 4317–4327, 10 2018. [Online]. Available: https://doi.org/10.1007/s00542-018-3769-6

[59] E. Digo, S. Pastorelli, and L. Gastaldi, "A Narrative Review on Wearable Inertial Sensors for Human Motion Tracking in Industrial Scenarios," *Robotics*, vol. 11, p. 138, 4 2022.

[60] E. Knippenberg, J. Verbrugghe, I. Lamers, S. Palmaers, A. Timmermans, and A. Spooren, "Markerless motion capture systems as training device in neurological rehabilitation: A systematic review of their use, application, target population and efficacy," *Journal of NeuroEngineering and Rehabilitation*, vol. 14, 4 2017.

[61] D. Laurijssen, S. Truijen, W. Saeys, W. Daems, and J. Steckel, "An Ultrasonic Six Degrees-of-Freedom Pose Estimation Sensor," *IEEE Sensors Journal*, vol. 17, no. 1, pp. 151–159, 2017.

[62] D. Laurijssen, S. Truijen, W. Saeys, and J. Steckel, "Three sources, three receivers, six degrees of freedom: An ultrasonic sensor for pose estimation & motion capture," in *2015 IEEE SENSORS*, 2015, pp. 1–4.

[63] D. Laurijssen, S. Truijen, W. Saeys, W. Daems, and J. Steckel, "A flexible embedded hardware platform supporting low-cost human pose estimation," in *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2016, pp. 1–8.

[64] Y. Qi, C. B. Soh, E. Gunawan, K.-S. Low, and R. Thomas, "Assessment of Foot Trajectory for Human Gait Phase Detection Using Wireless Ultrasonic Sensor Network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 88–97, 2016.

[65] C. Castellini, K. Hertkorn, M. Sagardia, D. S. González, and M. Nowak, "A virtual piano-playing environment for rehabilitation based upon ultrasound imaging," in *5th IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics*, 2014, pp. 548–554.

[66] S. Patwardhan, J. Schofield, W. M. Joiner, and S. Sikdar, "Sonomyography shows feasibility as a tool to quantify joint movement at the muscle level," in *2022 International Conference on Rehabilitation Robotics (ICORR)*, 2022, pp. 1–5.

[67] D. Weenk, D. Roetenberg, B.-J. J. F. van Beijnum, H. J. Hermens, and P. H. Veltink, "Ambulatory Estimation of Relative Foot Positions by Fusing Ultrasound and Inertial Sensor Data," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 5, pp. 817–826, 2015.

[68] S.-W. Seo and S. Kwon, "3D hand motion and position estimation using ultrasonic receiver array and inertial sensors," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 2021, pp. 287–289.

[69] W. Zhang, M. Tomizuka, and N. Byl, "A Wireless Human Motion Monitoring System for Smart Rehabilitation," *Journal of Dynamic Systems, Measurement, and Control*, vol. 138, no. 11, 7 2016. [Online]. Available: https://doi.org/10.1115/1.4033949

[70] V. D. Tsakanikas, D. Gatsios, D. Dimopoulos, A. Pardalis, M. Pavlou, M. B. Liston, and D. I. Fotiadis, "Evaluating the Performance of Balance Physiotherapy Exercises Using a Sensory Platform: The Basis for a Persuasive Balance Rehabilitation Virtual Coaching System," *Frontiers in Digital Health*, vol. 2, 2020. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2020.545885

[71] H. Xu, L. Gao, H. Zhao, H. Huang, Y. Wang, G. Chen, Y. Qin, N. Zhao, D. Xu, L. Duan, X. Li, S. Li, Z. Luo, W. Wang, and Y. Lu, "Stretchable

and anti-impact iontronic pressure sensor with an ultrabroad linear range for biophysical monitoring and deep learning-aided knee rehabilitation," *Microsystems & Nanoengineering*, vol. 7, no. 1, p. 92, 2021. [Online]. Available: https://doi.org/10.1038/s41378-021-00318-2

[72] S. Z. Gurbuz and M. G. Amin, "Radar-Based Human-Motion Recognition With Deep Learning: Promising Applications for Indoor Monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[73] P. A. Schooley and S. A. Hamza, "Radar Human Motion Classification Using Multi-Antenna System," 2021.

[74] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward Unobtrusive In-Home Gait Analysis Based on Radar Micro-Doppler Signatures," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2629–2640, 2019.

[75] D. G. Bresnahan and Y. Li, "Hip Motion Measurement and Classification Using Millimeter Wave Radar and Convolutional Neural Networks," in *2022 IEEE Texas Symposium on Wireless and Microwave Circuits and Systems (WMCS)*, 2022, pp. 1–5.

[76] T. Zheng, Z. Chen, S. Zhang, C. Cai, and J. Luo, "MoRe-Fi: Motion-Robust and Fine-Grained Respiration Monitoring via Deep-Learning UWB Radar," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 111–124. [Online]. Available: https://doi.org/10.1145/3485730.3485932

[77] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "Mm-Pose: Real-Time Human Skeletal Posture Estimation Using mmWave Radars and CNNs," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 9 2020.

[78] S. Yoon, H.-W. Jung, H. Jung, K. Kim, S.-K. Hong, H. Roh, and B.-M. Oh, "Development and Validation of 2D-LiDAR-Based Gait Analysis Instrument and Algorithm," *Sensors*, vol. 21, no. 2, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/2/414

[79] H. T. Duong and Y. S. Suh, "Human Gait Tracking for Normal People and Walker Users Using a 2D LiDAR," *IEEE Sensors Journal*, vol. 20, no. 11, pp. 6191–6199, 2020.

[80] A. K. Patil, A. Balasubramanyam, J. Y. Ryu, B. Chakravarthi, and Y. H. Chai, "An Open-Source Platform for Human Pose Estimation and Tracking Using a Heterogeneous Multi-Sensor System," *Sensors*, vol. 21, no. 7, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/7/2340

[81] A. K. Patil, A. Balasubramanyam, J. Y. Ryu, P. K. B N, B. Chakravarthi, and Y. H. Chai, "Fusion of Multiple Lidars and Inertial Sensors for the Real-Time Pose Tracking of Human Motion," *Sensors*, vol. 20, no. 18, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5342

[82] Y. Wang, L. Guo, Z. Lu, X. Wen, S. Zhou, and W. Meng, "From Point to Space: 3D Moving Human Pose Estimation Using Commodity WiFi," *IEEE Communications Letters*, vol. 25, no. 7, pp. 2235–2239, 2021.

[83] Y. Ren, Z. Wang, S. Tan, Y. Chen, and J. Yang, "Winect: 3D human pose tracking for free-form activity using commodity WiFi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, 12 2021.

[84] J. Geng, D. Huang, and F. la Torre, "DensePose From WiFi," 2023. [Online]. Available: https://arxiv.org/abs/2301.00250

[85] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, and J. Yang, "3D Human Pose Estimation Using WiFi Signals," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. ACM }, 11 2021.

[86] S. M. Hernandez, M. Touhiduzzaman, P. E. Pidcoe, and E. Bulut, "Wi-PT: Wireless Sensing based Low-cost Physical Rehabilitation Tracking," in *2022 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, 2022, pp. 113–118.

[87] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," 2022. [Online]. Available: https://arxiv.org/abs/2211.03375

[88] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[89] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity }Fields," *CoRR*, vol. abs/1812.08008, 2018. [Online]. Available: http://arxiv.org/abs/1812.08008

[90] S. Yang, Z. Quan, M. Nie, and W. Yang, "TransPose: Towards Explainable Human Pose Estimation by Transformer," *CoRR*, vol. abs/2012.14214, 2020. [Online]. Available: https://arxiv.org/abs/2012.14214

[91] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "TFPose: Direct Human Pose Estimation with Transformers," *CoRR*, vol. abs/2103.15320, 2021. [Online]. Available: https://arxiv.org/abs/2103.15320

[92] D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," *CoRR*, vol. abs/1811.12004, 2018. [Online]. Available: http://arxiv.org/abs/1811.12004

[93] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," *CoRR*, vol. abs/1705.03098, 2017. [Online]. Available: http://arxiv.org/abs/1705.03098

[94] Z. Zhang, J. Tang, and G. Wu, "Simple and Lightweight Human Pose Estimation," in *Computer Vision and Pattern Recognition*, 2020.

[95] H. Ren, W. Wang, K. Zhang, D. Wei, Y. Gao, and Y. Sun, "Fast and Lightweight Human Pose Estimation," *IEEE Access*, vol. 9, pp. 49 576–49 589, 2021.

[96] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," *CoRR*, vol. abs/1907.05047, 2019. [Online]. Available: http://arxiv.org/abs/1907.05047

[97] S. Mroz, N. Baddour, C. McGuirk, P. Juneau, A. Tu, K. Cheung, and E. Lemaire, "Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, 2021, pp. 1–4.

[98] A. Vakanski, J. Ferguson, and S. Lee, "Metrics for Performance Evaluation of Patient Exercises during Physical Therapy," *American Journal of Physical Medicine & Rehabilitation*, vol. 5, 1 2017.

[99] Y. Liao, A. Vakanski, and M. Xian, "A Deep Learning Framework for Assessing Physical Rehabilitation Exercises," *CoRR*, vol. abs/1901.10435, 2019. [Online]. Available: http://arxiv.org/abs/1901.10435

[100] A. Vakanski, H.-P. Jun, D. Paul, and R. Baker, "A Data Set of Human Body Movements for Physical Rehabilitation Exercises," *Data*, vol. 3, p. 2, 1 2018.

[101] N. Eichler, H. Hel-Or, I. Shmishoni, D. Itah, B. Gross, and S. Raz, "Non-Invasive Motion Analysis for Stroke Rehabilitation using off the Shelf 3D Sensors," 1 2018, pp. 1–8.

[102] A. R. Fugl-Meyer, L. Jääskö, I. A. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance." *Scandinavian journal of rehabilitation medicine*, vol. 7 1, pp. 13–31, 1975.

[103] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation In The Wild," *CoRR*, vol. abs/1802.00434, 2018. [Online]. Available: http://arxiv.org/abs/1802.00434

[104] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[105] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 7 2014.

[106] C. Mandery, Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The KIT whole-body human motion database," *2015 International Conference on Advanced Robotics (ICAR)*, pp. 329–336, 2015.

[107] G. Saeed and K. A. N. D. T. A. A. N. D. K. K. A. N. D. C. D. J. A. N. D. B. G. A. N. D. T. N. F. Mahdaviani, "MoVi: A large multi-purpose human motion and video dataset," *PLOS ONE*, vol. 16, no. 6, pp. 1–15, 1 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0253157

[108] E. Dolatabadi, Y. Zhi, B. Ye, G. Lupinacci, A. Mihailidis, R. Wang, and B. Taati, "The Toronto Rehab Stroke Pose Dataset to Detect Compensation during Stroke Rehabilitation Therapy." ACM, 1 2018.

[109] D. Leightley, M. H. Yap, J. Piasecki, Y. Barnouin, and J. Mcphee, "Benchmarking Human Motion Analysis Using Kinect One: an open source dataset," 1 2015.

[110] A. Miron, N. Sadawi, W. Ismail, H. Hussain, and C. Grosan, "IntelliRehabDS (IRDS)—A Dataset of Physical Rehabilitation Movements," *Data*, vol. 6, p. 46, 1 2021.

[111] M. Naeemabadi, B. Dinesen, O. K. Andersen, and J. Hansen, "Investigating the impact of a motion capture system on Microsoft Kinect v2 recordings: A caution for using the technologies together," *PLoS ONE*, vol. 13, no. 9, 9 2018.

[112] 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*, vol. 27, no. 7, pp. 1436–1448, 2019.

[113] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, S. Longhi, L. Romeo, S. N. Russi, and F. Verdini, "Accuracy evaluation of the Kinect v2 sensor during dynamic movements in a rehabilitation scenario," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 5409–5412.

[114] M. Capecci, M. G. Ceravolo, F. Ferracuti, M. Grugnetti, S. Iarlori, S. Longhi, L. Romeo, and F. Verdini, "An instrumental approach for monitoring physical exercises in a visual markerless scenario: A proof of concept," *Journal of Biomechanics*, vol. 69, pp. 70–80, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021929018300228

[115] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[116] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[117] Bassett Biomechanics, "Visual3D." [Online]. Available: https://bassettbiomechanics.com/visual3d/

[118] M. Capecci, M. Ceravolo, F. orazio, F. Ferracuti, S. Iarlori, G. Lazzaro, S. Longhi, L. Romeo, and F. Verdini, "A tool for home-based rehabilitation allowing for clinical evaluation in a visual markerless scenario," in *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, vol. 2015, 4 2015.

[119] geaxgx, "BlazePose Tracking with DepthAI." [Online]. Available: https://github.com/geaxgx/depthai_blazepose

[120] "Depthai's documentation¶." [Online]. Available: https://docs.luxonis.com/en/latest/

[121] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks
}via Gradient-based Localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391