1290

UNIVERSIDADE Đ
COIMBRA

Miguel Ângelo Soares Cerveira Varandas

3D Human Pose and Shape Estimation for Autonomous Telerehabilitation Systems

VOLUME 1

Dissertation in the Master's in Eletrical and Computer Engineering, supervised by Professor Dr. Paulo José Monteiro Peixoto and Professor Dr. João Luís Ruivo Carvalho Paulo and presented to the Faculty of Science and Technology of the University of Coimbra.

September 2023

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

3D Human Pose and Shape Estimation for Autonomous Telerehabilitation Systems

Miguel Ângelo Soares Cerveira Varandas

Coimbra, Junho 2023

3D Human Pose and Shape Estimation for Autonomous Telerehabilitation Systems

Supervisor:

Professor Dr. Paulo José Monteiro Peixoto

Co-Supervisor:

Professor Dr. João Luís Ruivo Carvalho Paulo

Jury:

Professor Dr. João Pedro de Almeida Barreto

Professor Dr. Paulo Jorge Carvalho Menezes

Dissertation submitted in partial fulfillment for the degree of Master of Science in Electrical and Computer Engineering.

Coimbra, Junho 2023

Acknowledgements

I would like to express my deepest appreciation and gratitude to the individuals who have supported and contributed to the completion of this dissertation. Their unwavering support, guidance, and the joyous moments we shared have made this journey truly memorable.

First and foremost, I extend my heartfelt thanks to ISR (Institute of Systems and Robotics of the Department of Electrical Engineering and Computers of the University of Coimbra) for providing me with a conducive academic environment and the necessary resources to pursue my research. The opportunities for growth, collaboration, and intellectual stimulation that this institute has offered have been instrumental in shaping my academic journey.

I am profoundly grateful to my two advisor professors, Professor Dr. Paulo José Monteiro Peixoto and Professor Dr. João Luís Ruivo Carvalho Paulo. Their expertise, guidance, and constructive feedback have been invaluable throughout the dissertation process. Their mentorship and encouragement have pushed me to explore new ideas and overcome challenges. Moreover, the laughter-filled conversations and the joyous moments we shared during our meetings have made this academic endeavour all the more enjoyable.

To my dear father and mother, I owe an immeasurable debt of gratitude. Your unwavering love, support, and belief in me have been my constant source of strength and inspiration. Your sacrifices, guidance, and encouragement have been the bedrock upon which I built my academic pursuits. The laughter and good times we shared as a family have provided much-needed respite and joy during the challenging phases of this journey. Thank you for always being there for me.

Finally, I would like to acknowledge my dear friend Pedro Miguel Cera Ramos dos Santos. Your friendship has brought immense joy and laughter to my life. Our shared moments of laughter, inside jokes, and the adventures we embarked upon have provided a refreshing break from the demands of academic life. Your unwavering support and ability to bring a smile to my face, even during the most stressful times, have been invaluable.

Resumo

Nas últimas duas décadas, assistiu-se a um aumento considerável de pacientes crónicos, incapacitados ou com mobilidade reduzida, devido principalmente ao aumento da esperança média de vida e da população mundial. Entre estes inúmeros casos, a dor lombar destaca-se por ser o sintoma que mais preocupação gera entre os profissionais de saúde, não só por ser capaz de afetar todas as faixas etárias, mas por ser bastante dispendiosa em termos de pagamento de planos de saúde, invalidez e absentismo no trabalho. A área da fisioterapia desempenha um papel fundamental na redução e prevenção da perda funcional desses pacientes por meio de exercícios e manutenção da atividade física. Contudo, certos fatores como a crescente demanda por este tipo de serviços, a consequente falta de recursos humanos e a pandemia do COVID-19 obrigaram a adoção de novas soluções como, por exemplo, a telereabilitação.

O projeto Intelligent Platform For Autonomous Collaborative Telereabilitation (INPACT) surge da ligação destas ideias, e visa o desenvolvimento de uma plataforma de telereabilitação de baixo custo, com uma interface de utilizador capaz de sugerir exercícios remotamente pré-configurados por um terapeuta. O sistema é capaz de monitorar o desempenho e movimento do corpo do utente por meio de uma câmara e fornecer feedback por meio de mecanismos de Deep Learning. O principal foco desta dissertação será assim a implementação do algoritmo responsável pelo processamento de um conjunto de imagens de vídeo, a estimação do esqueleto, da forma e da pose do indivíduo, e a criação de um modelo tridimensional virtual do corpo de cada paciente. A partir deste, será possível extrair os vértices da malha corporal obtida e segmentar a curvatura da coluna, que será posteriormente avaliada e comparada com a execução correta do exercício para que o utilizador possa ajustar a sua postura e evitar o risco de lesões.

O método utilizado, designado por HybrIK, consiste então numa solução inovadora de cinemática inversa neuro-analítica híbrida, responsável por encontrar as rotações relativas que permitem produzir as localizações desejadas das articulações do corpo. Por meio de uma decomposição *twist-and-swing*, cada parte do esqueleto é decomposta numa rotação longitudinal e uma rotação no plano, compostas ao longo da árvore cinemática, calculando a rotação de *swing* e prevendo a rotação de *twist*. Além disso, o seu alto desempenho advém também da adoção de mapas de calor volumétricos na representação final da aprendizagem das localizações das articulações 3D, estimados pela rede neuronal de alta resolução HRNet-W48, que permite associar as articulações 3D com o modelo paramétrico SMPL do corpo humano. Com a malha tridimensional do paciente resultante, recorreu-se a um conjunto de bibliotecas do Python para que, mediante um processo de interpolação, seja possível segmentar a curvatura da coluna.

Por fim, de modo a avaliar o HybrIK, foram também conduzidos alguns testes com os datasets Human3.6M, MPI-INF-3DHP, COCO e 3DPW, relativamente à sua robustez e capacidade de correção de erros, e uma pequena análise ao ângulo de twist estimado.

Keywords: Telereabilitação, Machine Learning, Deep Learning, Estimação Tridimensional Humana, Modelo Paramétrico, Interpolação, Curvatura da Coluna.

Abstract

Over the past two decades, a significant rise in chronic patients with disabilities or reduced mobility has been observed, primarily attributed to the increased global life expectancy and population growth. Among these cases, lower back pain stands out as a symptom of paramount concern to healthcare professionals. Its impact spans across all age groups and imposes substantial financial burdens in terms of healthcare payments, disability benefits, and work absenteeism. The field of physiotherapy assumes a pivotal role in mitigating and preventing functional decline in these patients through exercise and physical activity maintenance. However, factors like the escalating demand for such services, concurrent shortages in human resources, and the COVID-19 pandemic necessitated the adoption of innovative solutions, such as tele-rehabilitation.

The Intelligent Platform For Autonomous Collaborative Telereabilitation (INPACT) project emerges from these notions, aiming to create a cost-effective telerehabilitation platform. The platform integrates a user interface capable of remotely suggesting pre-configured exercises by a therapist. Through the utilization of deep learning mechanisms, the system monitors user performance and body movements via a camera, providing feedback. This dissertation primarily focuses on implementing an algorithm responsible for processing video images, estimating user skeletal structure, shape, and pose, and constructing a virtual 3D model of each patient's body. This model facilitates the extraction of vertices from the obtained body mesh, enabling curvature segmentation of the spine. This curvature assessment is then compared to the correct exercise execution, empowering users to adjust their posture and mitigate injury risk.

The proposed method, named HybrIK, introduces an innovative solution through a hybrid neuro-analytical inverse kinematics approach. It determines the relative rotations necessary to achieve desired joint positions within the body. Leveraging a twist-and-swing decomposition, each skeleton segment is broken down into a longitudinal and a plane-based rotation, compounded along the kinematic tree. Swing and twist rotations are then calculated and predicted, respectively. The method's efficiency is further enhanced by the integration of volumetric heatmaps in the final representation of 3D joint locations. These are estimated using the high-resolution neural network HRNet-W48, allowing association with the parametric SMPL model of the human body. The resultant 3D patient mesh is processed through Python libraries, employing interpolation to segment spinal curvature.

Ultimately, to assess HybrIK's performance, tests were conducted with the Human3.6M, MPI-INF-3DHP, COCO, and 3DPW datasets. These tests evaluated the method's robustness, error correction capabilities, and provided an analysis of the estimated twist angle.

Keywords: Tele-rehabilitation, Machine Learning, Deep Learning, Three-Dimensional Human Estimation, Parametric Model, Interpolation, Spinal Curvature.

“You may never know what results come of your actions, but if you do nothing, there will be no results."

— Mahatma Gandhi

Contents

List of Acronyms

AI	Inteligência Artificial
DNN	Deep Neural Network
DoFs	Degrees of Freedom
DYNA	Dynamic Human Shape in Motion
FK	Forward Kinematics
HPE	Human Pose Estimation
IK	Inverse Kinematics
INPACT	Intelligent Platform For Autonomous Collaborative Telereabilitation
ML	Machine Learning
PCA	Principal Component Analysis
SCAPE	Shape Completion and Animation of People
SMPL	Skinned Multi-Person Linear

List of Figures

List of Tables

1 Introduction

In the last two decades, the number of chronically ill, disabled or mobility-impaired people has risen sharply, mainly due to the increase in average life expectancy and thus in the world's population.

Low back pain is one of the problems of greatest concern to health professionals, not only because it has increased by about 50% since the 1990s, especially in less developed countries, but also because it can affect all age groups and is associated with sedentary occupations, smoking and certain diseases such as obesity [1]. In addition, there are patients who need post-operative care to avoid long hospital stays and hospital congestion.

Physiotherapists play a fundamental role in reducing and preventing the decline in physiological and functional capacity of these patients through appropriate training exercises and maintaining a high level of physical activity. Due to the increasing demand for this type of service, which requires careful and specific attention to each pathology and patient [2], [3], the resulting lack of human resources and the crisis caused by the COVID-19 pandemic, which forced the introduction of measures such as the reduction of face-to-face activities and the introduction of an online or remote consultation modality, health systems are currently open to new innovation processes to improve the effectiveness and efficiency of the health services provided [2].

Telerehabilitation is one such development that enables the treatment of acute phases of illness by replacing the traditional face-to-face approach in the interaction between patient and physiotherapist. Although it is a relatively new field of research, it covers situations where it becomes difficult for patients to travel to rehabilitation facilities, which are often far from their homes. Consequently, studies suggest that clients are more likely to attend training sessions because of the convenience of being at home, and that healthcare professionals can better control the time, intensity and sequence of the intervention. Ultimately, this leads to environmental benefits as clients are less likely to travel [4]. Thus, the term telerehabilitation can be defined as the provision of rehabilitation services to patients remotely, usually through the use of virtual reality, augmented reality, motion capture technologies and, in more complex cases, systems based on machine learning [5]. One example is the so-called *exergames*, which aim to stimulate the movement of the patient's body through virtual and interactive environments that simulate different sensations and require some physical effort to play.

The exponential development of certain areas such as the internet, mobile devices and artificial intelligence means that huge applications and services that put people at the centre are playing an increasingly important role, because body language is one of the simplest forms of communication. Considering this, it is easy to see that a computer that better understands the movements of the human body will lead to significant advances and the building of a better system of interaction between the two.

1.1 Motivation

One of the most interesting and complex challenges in the fields of computer vision and artificial intelligence is the analysis and recognition of humans on images, which is divided into tasks such as action recognition, 3D reconstruction, segmentation and, what arouses the most interest in the context of this dissertation, the three-dimensional study and estimation of the posture and shape of the human body.

Since 3D estimation provides additional depth information compared to 2D estimation, it is suitable for a variety of real-world applications [6]:

- **Movies and video games:** in the film industry, special effects are widely known to allow the three-dimensional simulation of actors' human bodies to produce scenes that are considered dangerous or unattainable in reality, such as science fiction films set in outer space. In video games, an accurate and realistic human body becomes a crucial factor in providing the best experience and entertainment.

- **Fashion:** due to the growing popularity of online shopping, driven by the COVID-19 pandemic, ordering clothes and even shoes became more accessible with the creation of a virtual testing system based on the customer's three-dimensional assessment. Users can try on different sizes and styles, avoiding complications and delays in returns, for example.

- **Autonomous driving:** when we consider an autonomous driving vehicle, it is easy to

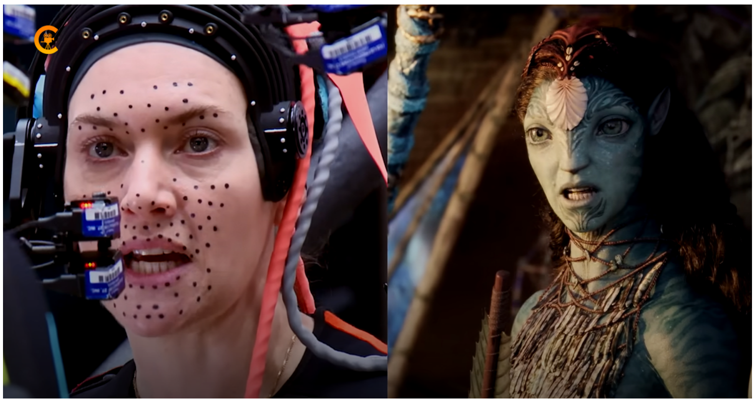


Figure 1.1: Example of pose estimation and motion capture on the set of "Avatar: The Way of Water".

understand that it has to make decisions to avoid collisions with pedestrians. Therefore, it is important to understand its attitude, movement and intention in real time.

- **Video surveillance:** video surveillance is of great importance to public safety as it is present in certain services from banks and large shopping centres to traditional shops and petrol pumps. In this area, body posture estimation techniques can help surveillance officers detect and identify suspicious persons.
- **Health and physiotherapy:** a patient's posture and movement can indicate his state of health. In this way, the three-dimensional assessment of pose allows healthcare professionals to remotely diagnose, monitor the execution of certain rehabilitation exercises and thus provide feedback to correct posture and prevent injury on the part of the patient.

1.2 Challenges

Despite its wide application, the topic of three-dimensional estimation of human posture has become quite popular in the scientific community, mainly because, unlike two-dimensional estimation, it presents some unique challenges, including the unknown position of the person in the image, the possible presence of multiple people in the scene, the heavy occlusion of body elements, the lack of in-the-wild data, a huge variety of actions and body shapes, clothing, and much more [6], [7].

1. **Several persons:** in the case of monocular images, estimating the pose of multiple people compared to a single person becomes a much more complex task, solely due to

occlusion by people who are close together, as shown in the figure 1.2. In the context of video, a new challenge arises when considering the possibility of different angles of the scene. This includes the ambiguity of cross-viewing, which involves geometric constraints due to the overlapping of fields of view [8], [9].

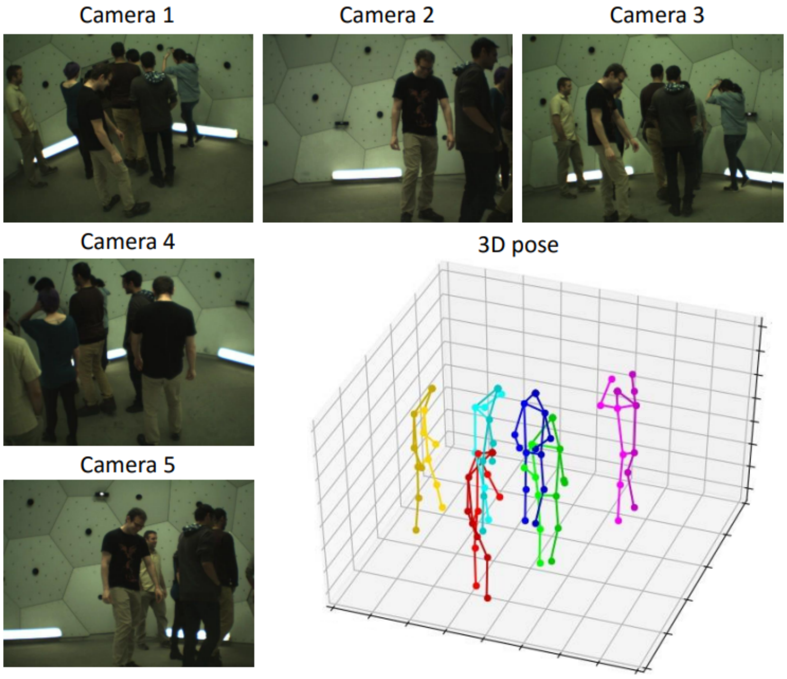


Figure 1.2: Illustrative example of a model whose aim is to remove the three-dimensional pose of several people from different viewpoints, with problems such as occlusion and overlapping fields of vision [10].

2. **Occlusion:** the term appears in situations where certain parts of the person's body cannot be directly observed by the camera because they are covered by objects, which is called occlusion, but also in cases where the person performs certain types of actions or is not properly oriented, as self-occlusion. Since a person's pose is usually represented by the position of the points of articulation, there is a great deal of uncertainty and freedom in estimating them if they are not directly identified in the image.

3. **In-the-wild data:** in the case of two-dimensional estimation, the creation of large datasets in nature becomes possible because the pose of the human can be easily labelled manually. In the case of three-dimensional annotations, the situation is quite different, as they are captured by proprietary marker-based image processing systems, which takes a lot of resources and time when creating large datasets. Furthermore, for different scenarios, there are factors such as lighting, shadows and background colours

that can add noise and negatively affect the segmentation of the human body for the many existing algorithms.

4. **Depth ambiguity:** as one of the main problems of three-dimensional estimation of static images, it arises from the inversion of a nonlinear lossy transformation that combines kinematics with perspective projection. Due to the fact that the relative depth between the joints of the human body is unknown, which means that several 3D poses can correspond to a single 2D pose, this leads to depth ambiguities between the two-dimensional and the three-dimensional projection.

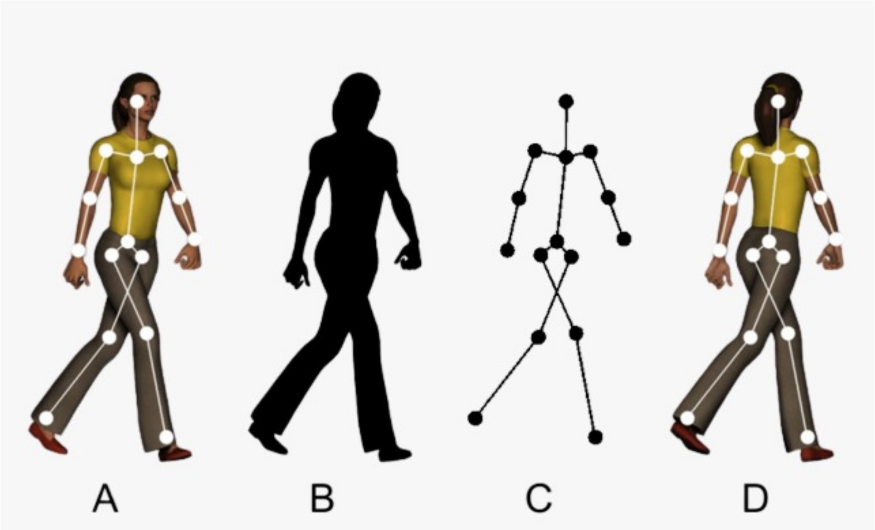


Figure 1.3: Illustrative example of depth ambiguity.

To make it clearer, consider the figure 1.3 where the silhouette B and the skeleton C correspond exactly to the two poses represented by A and D. The stimuli shown in B and C can be interpreted as if the person were looking forward or backward, since the depth order of the joints is ambiguous in each representation.

1.3 Problem statement

Combining these ideas, the Intelligent Platform For Autonomous Collaborative Telerehabilitation (INPACT) project is born, which aims to develop a low-cost telerehabilitation platform with a user interface capable of remotely suggesting pre-configured exercises related to lower back pain by a therapist. The system will be able to monitor the performance and movement of the user's body via a camera, without having to resort to the usual markers, thus providing feedback via ML mechanisms.

The focus of the present work is to develop an algorithm that processes a series of video images and then estimates the person's skeleton, shape and posture to create a virtual and animated three-dimensional model of each patient's body. From this, the vertices of the body mesh obtained can be extracted and the curvature of the spine estimated. This is later evaluated and compared with the correct execution of the exercise previously specified by the health expert. In this way, the user can adjust their posture and perform the different exercises correctly, avoiding the risk of injury.

1.4 System Goals

As previously mentioned, the proposed system will employ a Machine Learning-based approach capable of processing a large set of images depicting the execution of various rehabilitation exercises recommended by a physiotherapist. Thus, the objectives of this thesis are appropriately presented:

- Research, study, and testing of an existing three-dimensional human estimation model.
- Processing and analysis of video frames from exercise executions.
- Estimation of the patient's pose and shape using the proposed model.
- Creation of a volumetric mesh of the patient.
- Error calculation and prediction assessment.
- Segmentation of the patient's spinal curvature.

1.5 Thesis outline

The current document is organized into 6 chapters. Chapter 1 serves as an introduction, providing context to the problem, its applications, associated challenges, and a final solution to mitigate it.

Chapter 2 discusses the essential concepts of Deep Learning networks related to Computer Vision and their evolution into the implemented High-Resolution Network. It also includes a brief overview of 2D and 3D human pose estimation methods and the three existing representations of the human body.

Moving on to Chapter 3, following the structure of the previous chapter, it presents and analyses some of the existing human pose estimation solutions and systems found in the literature, particularly those related to telerehabilitation.

Chapter 4 provides a detailed explanation of the employed models, specifically the human representation model SMPL based on 3D meshes and the human shape and pose estimation method called HybrIK. This chapter encompasses all the necessary considerations for its development, theoretical foundations, mathematical calculations, and expected benefits.

In Chapter 5, the datasets used for training and testing the network are analysed, and a series of tests are conducted to assess the performance of the HybrIK method. These tests specifically focus on inverse kinematics concepts, twist-and-swing decomposition, and its error correction capabilities. Additionally, the patient's spine segmentation process during telerehabilitation exercises is thoroughly explained.

Finally, Chapter 6 presents the final conclusions along with some points for potential future work.

2 Background

This chapter presents the fundamental principles of Deep Learning networks in the context of Computer Vision, tracing their development to the established High-Resolution Network implementation. Additionally, a concise survey of techniques for 2D and 3D human pose estimation is provided, along with an examination of the three prevailing models representing the human body.

2.1 Machine Learning

The last decade has seen a massive increase in the use of artificial intelligence (AI) for various applications, especially the aforementioned ML technology. Although these two terms are commonly confused, ML is only a small branch of AI responsible for the development of algorithms aimed at giving a machine the ability to "learn", i.e. to improve its performance on certain tasks based on previous experience or previously provided data. Consequently, ML relies on the availability of data to train the machine to perform the desired tasks [11].

By its nature, it is a method suitable for applications where the input data is used to produce an output based on some features of the same inputs, e.g. classification and image processing. In this context, it is easy to see that ML is one of the most interesting and promising fields for medical research applications [11].

2.1.1 Convolutional Neural Networks

A convolutional neural network, or CNN, is a subset of ML and is characterised by being the centre of Deep Learning algorithms. Its architecture is analogous to the connection patterns of neurons in the human brain and was strongly inspired by the organisation of the visual cortex. For example, when we are confronted with an image, our brain processes an enormous amount of information in a very short time, each neuron being in its specific receptive field and connected to the others to cover the entire visual field. Just as each neuron responds to this type of stimulus, the neurons in CNNs will only process data in their receptive field. Networks thus consist of several layers of nodes: an initial input layer, one or more hidden layers and finally an output layer. Each of these nodes is connected to another and has a corresponding weighting and limitation. When the output of a single node is above this set threshold, the node is activated and sends the data to the next layer of the network. Otherwise, no data will be shared. Their organisation thus recognises simpler patterns such as lines and curves at an early stage and relatively complex patterns, for example faces or objects, at a later stage.

But what is the function of the layers of a Convolutional Neural Network? CNNs are distinguished from other neural networks by their high performance, regardless of whether the input is an image, speech or audio signal, and are divided into three main types of layers.

Convolution Layers

The convolutional layer corresponds to the central building block of a CNN, where most of the computation takes place. It usually consists of a few components, divided into input data, a filter and a feature map. Let us say that your input is a colour image consisting of a 3D array of pixels. This means that it consists of three dimensions - height, width and depth - corresponding to their RGB values. The convolution process thus implies that a feature detector, commonly called a filter or kernel, goes through each of the receptive fields of the image and checks whether a particular feature is present.

More specifically, the kernel consists of a two-dimensional matrix of weights, usually 3x3 in size, which, when applied to a small area of the image, allows the scalar product between the input pixels and their weights to be calculated, and these results are stored in an output array. Next, the filter goes one step further and repeats the process until the kernel has gone through the entire image. Also, after each convolution operation, an activation function is applied to the resulting map to add non-linearity to the model. The final output of the scalar products is then called the feature or activation map, Figure 2.1.

A CNN is not limited to only one layer of this type, i.e. a second convolutional layer may follow the first. In these cases, the structure of the CNN becomes hierarchical, as the subsequent layers have access to the pixels of the receptive fields of the previous layers. As an example, suppose we are trying to determine whether a particular image contains a bicycle. If we consider the latter as a sum of parts, consisting of frame, handlebars, wheels and pedals, each individual part will form a pattern at a lower level of the network. And the combination of all the parts represents a higher level, creating a hierarchy of features in the CNN itself.

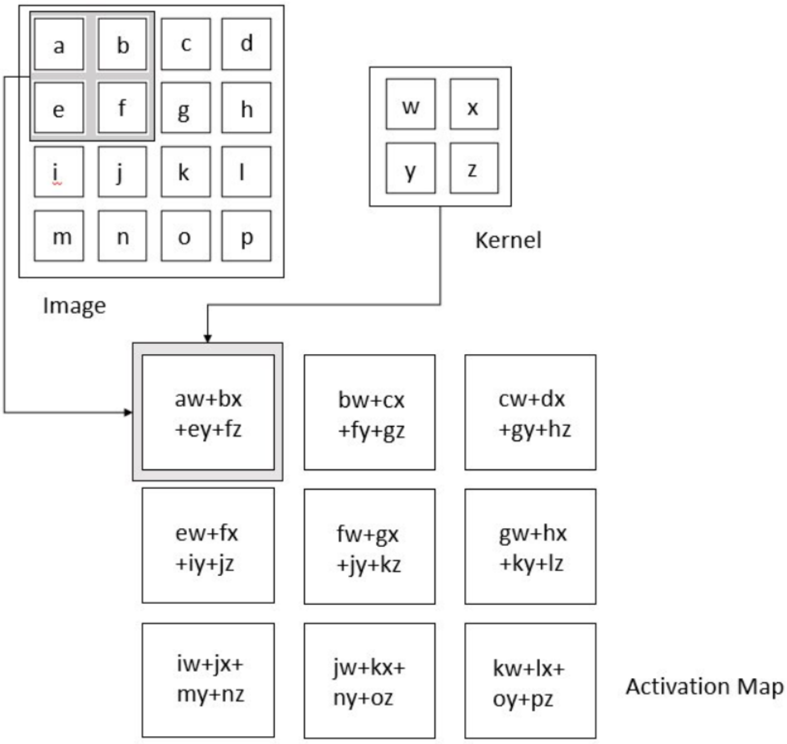


Figure 2.1: Convolution operation.

Pooling Layers

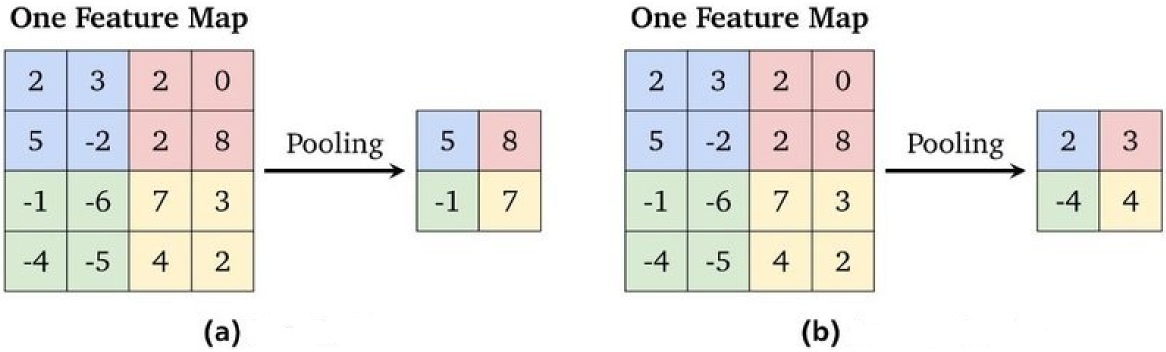
However, a limitation of the activation map output of convolutional layers is that they record the exact position of features in the input. This means that small changes in the position of a particular feature in the image will result in a different activation map, whether by cropping, rotating, shifting or other small changes. To solve this problem, pooling layers come into play, where lower-resolution versions of the input signals are created, consequently leading to a reduction in the computational power required to process them.

The pooling operation scans a second kernel along the activation map, except that in this case, there are no weights involved. Instead, the filter applies an aggregation function to the values of the receptive field, filling a new output matrix and extracting dominant features that are rotationally or positionally invariant, thus maintaining correct training of the model.

There are two main types of pooling:

- **Average pooling:** As the filter moves over the input, it calculates the average value within the receptive field to send to the output matrix.

- **Max pooling:** As the filter moves, it selects the pixel with the maximum value to send to the output matrix. Compared to the previous approach, this method tends to be used more frequently.



2	3	2	0
5	-2	2	8
-1	-6	7	3
-4	-5	4	2

(a) output:

5	8
-1	7

(b) output:

2	3
-4	4

Figure 2.2: (a) Max pooling. (b) Average pooling.

Fully-connected Layers

Finally, the fully connected layers within a Convolutional Neural Network (CNN) should not be confused with conventional fully connected neural networks. The conventional neural network architecture connects every neuron to every neuron in the subsequent layer. However, this traditional architecture was found to be inefficient for computer vision tasks. Images provide a substantial input to neural networks, potentially consisting of hundreds or even thousands of pixels and up to three colour channels. In a traditional fully connected network, accommodating such inputs necessitates an enormous number of connections and network parameters. To address this challenge, fully connected layers are typically found in the middle and/or at the end of neural network architectures. They take the output from convolutional and pooling layers and make predictions about the most suitable label to describe the image.

Until recently, this type of network enabled the construction of a new network, the HRNet, which forms the basis of the method for the practical part of this dissertation.

2.1.2 HRNet, High Resolution Network

With AlexNet [12], proposed in 2012, there was a rapid development in the architecture of CNNs in the field of computer vision, which included examples such as GoogleNet, VGGNet, Resnet and even DenseNet, as illustrated by Figure 2.3, used essentially for classification of images.

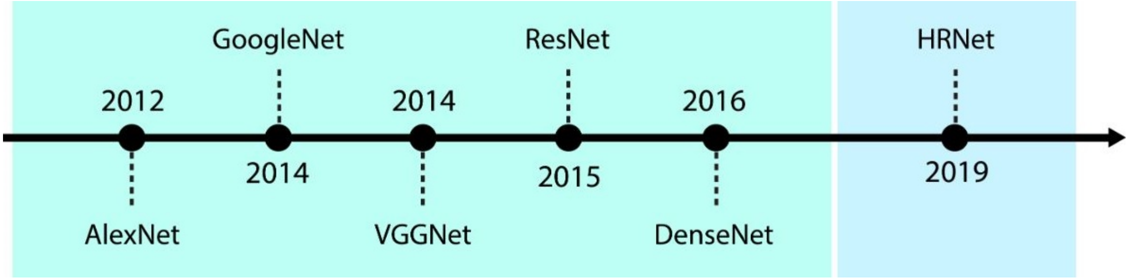


Figure 2.3: Development of networks in the context of computer vision (2012 - today).

However, tasks such as semantic segmentation, object detection and human pose estimation required spatially fine representations. Therefore, these networks started to extend the already existing classification architectures with high to low resolution subnetworks connected in series. And in the end, they increased the resolution.

But would it be possible to design a new universal architecture suitable for general computer vision tasks without considering classification tasks? The answer was given by a group of researchers with the development of the High Resolution Network. The name derives from the high resolution of the images to be processed and led to breaking the prevailing design rule by projecting it from scratch.

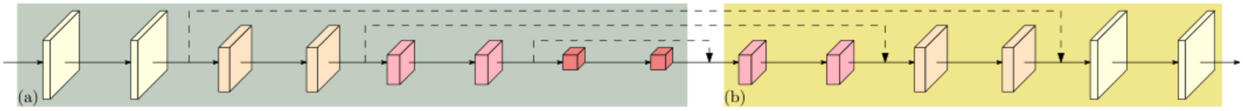


Figure 2.4: The structure of recovering high resolution from low resolution. (a) A low-resolution representation learning subnetwork (such as AlexNet, GoogleNet, VGGNet, ResNet, DenseNet), which is formed by connecting high-to-low convolutions in series. (b) A high-resolution representation recovering subnetwork, which is formed by connecting low-to-high convolutions in series. Representative examples include SegNet, DeconvNet, U-Net and Hourglass, and SimpleBaseline.

HRNet maintains these high-resolution representations throughout the whole process. The network starts from a high-resolution convolution stream, gradually adds high-to-low resolution convolution streams one by one, and connects the multi-resolution streams in parallel. The result thus consists of several stages, shown in blue in the Figure 2.5, where stage n contains n streams corresponding to n resolutions. It is then possible to perform repeated fusions with multiple resolutions by systematically exchanging information across the parallel streams.

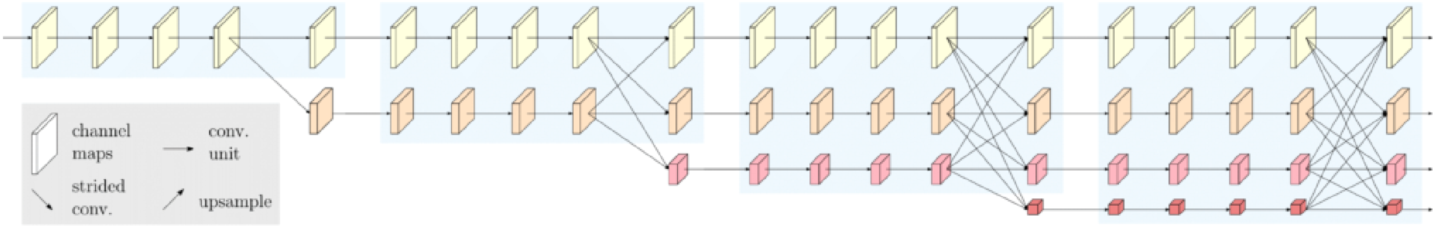


Figure 2.5: An example HRNet. There are four stages. The 1st stage consists of high-resolution convolutions. The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks several (that is, 1, 4, 3) times.

The high-resolution representations learned from HRNet are strongly spatially accurate. This is due to two aspects. First, because it is an approach that combines the high and low resolution convolutional streams in parallel rather than serially, it is able to retain the high resolution rather than recover it from the low resolution, allowing for a more accurate spatially learned representation. Second, although most existing fusion schemes aggregate high-resolution low-level and upsampled low-resolution high-level representations, in the case of HRNet, multi-resolution fusions are repeated to expand the high-resolution representations using the low-resolution representations through upsampling, and vice versa.

Since its release in 2019 [13], HRNet has become a standard network for human pose estimation, and there are signs in the literature that this is just the beginning of new and even more complex architectures.

2.2 Human Body Pose Estimation

The estimation of human posture is similar to the recognition of facial features, with the only exception that it is applied to the whole body and is more concerned with movement. Its main objective is to extract the posture of the human body from one or more images, by predicting and connecting body segments. Using the variety of methods and approaches for 2D and 3D HPE that have been introduced in recent years, the taxonomy shown in the Figure 2.6 can be created.

Starting on the left side of the tree, 2D HPE can first be divided into two upper levels of estimation. The SPPE methods, which are responsible for extracting the pose of only one person per image, and which are divided into approaches based on regression or body part detection; and MPPE, which on the other hand allow the extraction of the pose of multiple persons and can be divided into top-down or bottom-up approaches. In the right-hand branch, more specifically in 3D HPE, it is possible to observe two additional levels in terms of the type of input data for the model. The first case, Single-view, refers to images taken with only one camera. Multi-view, as the name suggests, indicates the type of images taken with multiple cameras placed at different angles. The topics coloured green in the last line of the tree, which require a specific and somewhat more extensive definition, can be found in the next chapter, accompanied by some of the most important examples from the literature.

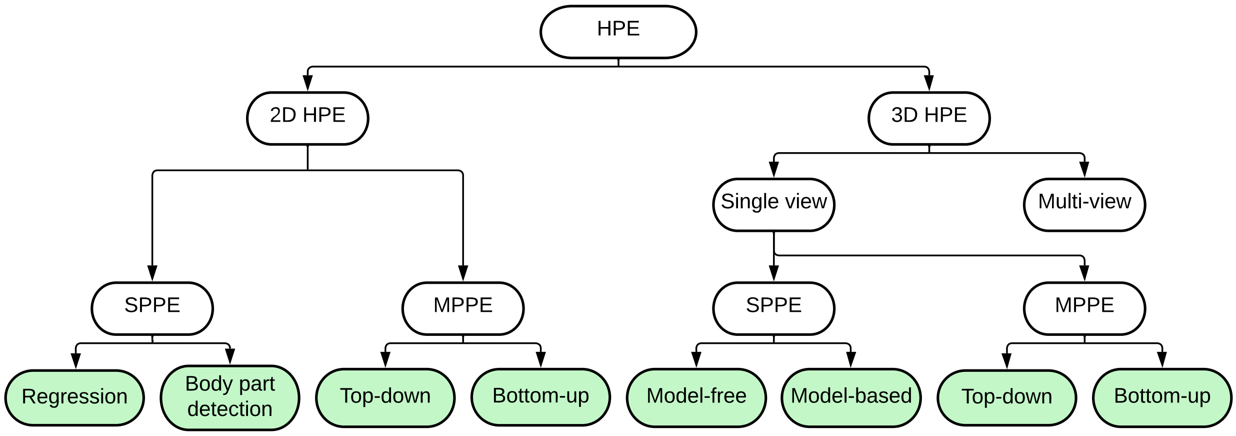


Figure 2.6: Taxonomy of HPE method types.

2.3 Human Body Modeling

Before we explain the above methods in more detail, it is important to introduce the different types of human representation used in each method.

Modelling the human body is undoubtedly one of the most important aspects in the HPE field, as it allows reproducing the key points and the different features extracted from the input data of a given model. However, it is easy to see that the human body is characterised by a highly complex and non-rigid kinematic structure, consisting of multiple joints and limbs and information about each person's particular morphology [14], which ultimately complicates the task of estimating and representing it. Nevertheless, and taking into account the different application scenarios, it is possible to divide the latter into three different types: Skeleton-based model, Contour-based model and Volume-based model [15].

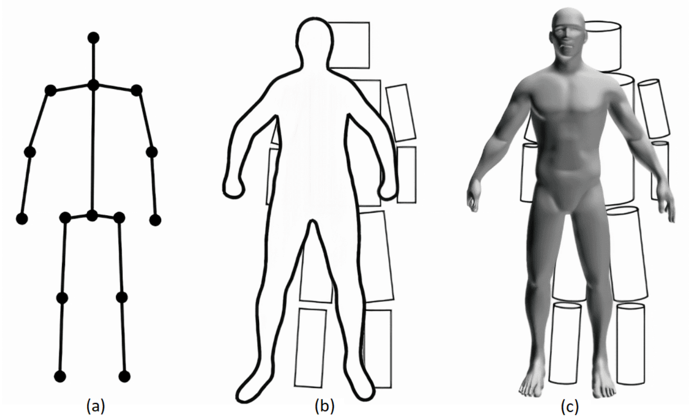


Figure 2.7: (a) Skeleton-based model. (b) Contour-based model. (c) Volume-based model.

2.3.1 Skeleton-based model

The first model, also called the kinematic model, is usually represented by a series of joint positions and the corresponding alignments of the individual limbs, which naturally follow the skeletal structure of the human body. The Figure 2.7.(a) shows one of these examples, which is usually described as a graph of edges connected by a set of nodes that form the respective limbs and joints. However, despite its simple and flexible representation, it has some limitations when it comes to body texture and contour information, so it is essentially used for 2D estimation [14], [15].

2.3.2 Contour-based model

The following contour-based model not only captures the relationships between the different parts of the body, but also contains approximate width and contour information for the limbs and torso [14], [15]. In this planar model used in the first methods of HPE, this problem was solved with small rectangles or boundaries of the person's silhouette, as shown in the Figure 2.7.(b).

2.3.3 Volume-based model

Finally, there is the volumetric model, which deserves special attention since it is the subject of study in this dissertation. The first models of this kind were based on the idea of the planar model and began by defining the various parts of the body by geometrical solids that comprised cylinders or cones. But with the recent advent of techniques based on Deep Learning, this kind of model has taken on a new representation in the form of a mesh, Figure 2.7.(c), made up of a huge set of vertices and faces usually captured by 3D scans of the human body [14]. Some of the most commonly used examples today are therefore *Shape Completion and Animation of People* (SCAPE), *Dynamic Human Shape in Motion* (DYNA) and *Skinned Multi-Person Linear* (SMPL).

SCAPE

In this first example, the model parameters are estimated using a sparse set of markers, namely 56, attached to the human body. The positions of these markers in space are later determined using a conventional motion capture system to obtain certain constraints on the shape of the body. In this way, the pose and shape parameters can be estimated so that the reconstructed body is forced to lie within the boundaries defined by the markers. All this presupposes that 3D scans of the subject are already available, since the said markers are to be placed there [16], [17].

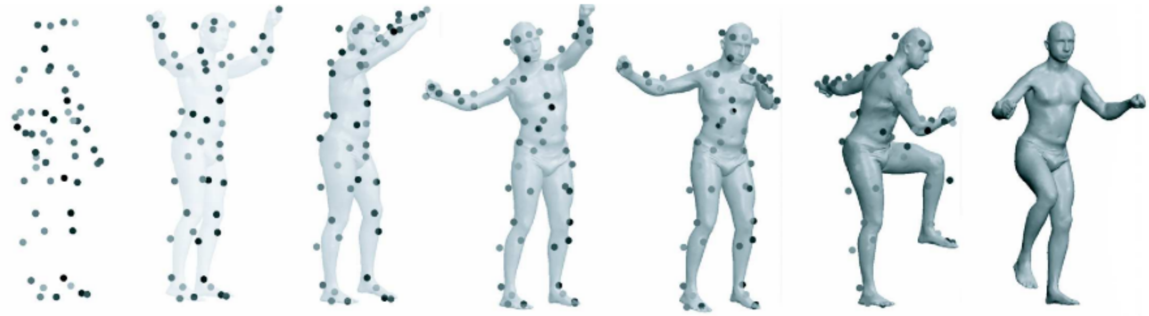


Figure 2.8: Application of the SCAPE model to the motion capture of a person of whom only one scan is available [16].

DYNA

The DYNA model stands out as one of the most complete in terms of manipulation and representation of the human body, paying particular attention to details such as the deformation of the soft tissues of a real human. Using a high-resolution 4D acquisition system and about $40,000$ scans of ten people, the relationship between soft tissue movement and deformation is approximated by a low-dimensional linear subspace. To this end, factors such as the velocity and acceleration of the whole body, the angular velocities and accelerations of the limbs and the shape coefficients of the soft tissues are considered to predict the low-dimensional coefficients. Finally, to make the model as general as possible with regard to the morphology of the human body, the body mass index of the person is also taken into account [18].

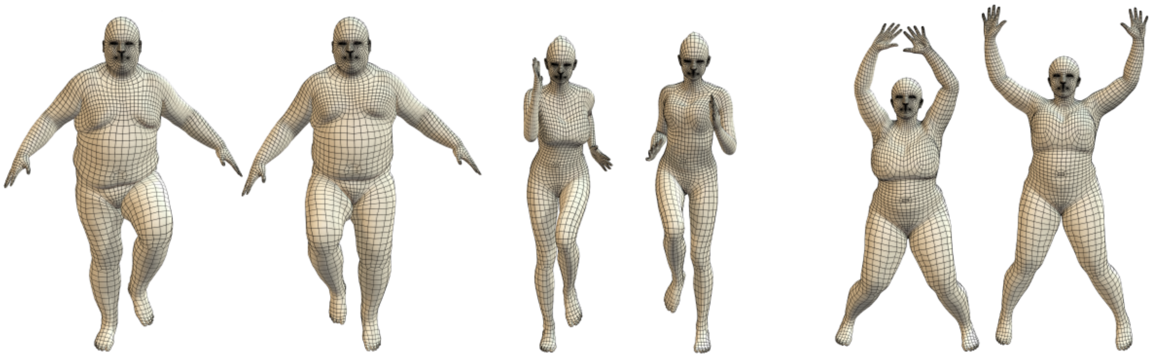


Figure 2.9: Three different animated body physiognomies performing a series of movements with the soft tissue deformations predicted by Dyna [18].

SMPL

In the case of the third model, the SMPL, the authors have resorted to the use of vertices to represent with high precision a variety of body shapes in natural human poses. Learning its parameters takes into account several types of data, divided into a model of the resting pose, weight mixtures, pose-dependent blending shapes and even a regression capable of obtaining the position of the joints from the vertices.

Unlike existing models, SMPL's pose-dependent blending forms differ in the sense that they become a linear function of the elements of the pose rotation matrices, which allows the model to be trained on thousands of three-dimensional meshes of different individuals in different positions. Furthermore, this model is not only compatible with any existing rendering engine, but it also applies the idea of DYNA and is thus able to realistically model soft tissue deformations [19].

The SMPL is the last example in the list of volumetric models for the simple reason that it was used in the implementation of the algorithm in this thesis. To make everything clearer, a detailed explanation of the method is given in chapter 4, including an illustration and the mathematical formulation used.

3 Related work

In recent years, the field of computer vision has witnessed remarkable advancements in the realm of 3D human pose and shape estimation. As the demand for accurate and robust techniques in understanding human body movements and shapes continues to grow, researchers have been actively exploring innovative approaches to tackle this challenging problem.

This chapter delves into the state-of-the-art methods and techniques employed in 2D and 3D human pose and shape estimation. It provides a comprehensive overview of the most recent advancements, highlighting both the achievements and the remaining challenges in this fascinating field.

3.1 2D Single Person Pose Estimation

The 2D SPPE aims to localise the position of the human body joints of a single person for each input image. If there is more than one person in an image, the image is cropped to show only one of them in each sub-image so that the SPPE method can move between them individually. Normally, this process is done automatically by applying detectors to the whole body [20] or the top half [21]. As mentioned earlier, 2D SPPE can be divided into two different categories due to the different formulations of the estimation task: Regression or Body Part Detection, which are described in the following subsections.

3.1.1 Regression

Regression methods use an end-to-end framework with the aim of automatically learning the association of input images with joints or body model parameters. The first method, developed by Toshev and Szegedy and known as DeepPose [22], is based on a Deep Neural Network developed by Krizhevsky *et al.* called AlexNet [12], and consists of a cascaded deep neural network capable of identifying key points in the input images. Its performance was so impressive that the HPE paradigm moved from classical approaches to Deep Learning, namely CNNs. Similarly, based on the existing GoogLeNet network [23], Carreira *et al.* [24] have proposed a new network, Iterative Error Feedback, which consists of a self-correcting model that incrementally changes an initial solution by feeding back the prediction error in the input space. According to them, their network is much more accurate and involves 12x fewer parameters than the architecture of Krizhevsky *et al.*

However, a critical issue with regression-based methods is the capture of features that encode important pose information. In this context, multitask learning became a very popular strategy for representing features. By sharing representations between correlated tasks, such as pose estimation and action recognition, the model could be better generalised in its original task, pose estimation. Li *et al.* [25] implemented a heterogeneous structure of this type that was responsible for predicting the coordinates of each joint from complete images and recognising body parts from a sliding window.

3.1.2 Body Part Detection

In this type of HPE technique, the aim is to train a body part detector to then estimate the position of its joints, which usually involves two processes. First, heat maps are made for the key points and then these estimated points are put together to form body postures. Specifically, this involves estimating K heat maps, $\{H_1, H_2, ..., H_K\}$, for a total of K key points, where the pixel value in each map indicates the probability that the key point is at position (x, y). In this way, the estimation networks are trained to minimise the discrepancy between the set of predicted heatmaps and the actual heatmaps.

Compared to joint coordinates, heatmaps provide richer monitoring information, as the preservation of their spatial position facilitates the training of convolutional networks. Wei *et al.* [26] began by introducing a sequential structure based on a set of convolutional networks, Convolutional Pose Machines, to predict the positions of key joints with multi-stage processing. This means that at each stage, the networks used the 2D maps created in the previous stages and gradually produced more refined predictions of the positions of the body parts.

In addition to efforts to design effective networks for this type of 2D HPE, the structure of the body is also being studied to provide more and better information for constructing these appropriate networks. Tang *et al.* [27] constructed a supervised hourglass-like network, the Deeply Learned Compositional Model, which describes the complex and realistic relationships between the different body parts and learns their compositional information such as orientation, scale and shape. Tang and Wu [28] realised that not all parts of the body are related when they introduced the Part-based Branches Network, whose task was to learn the specific representations for each group of parts.

3.2 2D Multi-Person Pose Estimation

Compared to individual HPE, this type of estimation is much more difficult and demanding, not only because you need to determine the number of individuals in the image and their positions, but also because you need to group the key points of each of them. There are two different methods to solve these problems: Top-Down and Bottom-Up.

3.2.1 Top-Down Approach

The first stage of this type of approach uses a number of standard segmentation methods, such as Faster R-CNN or Mask R-CNN, which aim to delineate each person in the image with a small box, the bounding boxes. In a next step, this allows an individual estimation of the pose of each of the persons identified with the SPPE techniques of the previous subchapter by connecting the key points and creating a 2D representation of the human body. This means that their calculation times are strongly influenced by the number of people in the input image.

To answer the question *"How good can a simple method be?"*, Xiao *et al.* [29] first added some layers of deconvolution to their ResNet support network to create a simple but effective structure for building heat maps for high-resolution representations. Sun *et al.* [13], on the other hand, built the HRNet network, as explained earlier, which learned reliable high-resolution representations by connecting subnetworks with multiple resolutions in parallel and repeatedly combining different scales.

However, in multi-person environments, occlusions and cut-offs often occur as members inevitably overlap, causing the person detectors used in the first stage to fail. Robustness to occlusions and cut-offs is therefore an essential aspect of this type of approach. With this in mind, Fang *et al.* [30] developed a multi-person pose estimation technique, RMPE, to improve performance in this type of complex scenario. Specifically, the RMPE structure was divided into three parts: a symmetric spatial transformer network that detected the region of a single subject within an imprecise bounding box, a parametric non-maximal pose suppression that solved the redundant detection problem, and a pose-guided proposal generator that improved data training.

3.2.2 Bottom-Up Approach

Bottom-up approaches first identify key points in each image, which are later grouped into individual subjects and combined to form the predicted poses. As can be readily seen, this leads to much shorter computation times compared to top-down methods, as the isolated identification of key points for each subject is not required.

Pishchulin *et al.* [31] proposed DeepCut, a body part detector based on Fast R-CNN. It first detected all the body parts in question and then labeled and assembled each of these parts into a final pose using integer linear programming. But the DeepCut model had one flaw: it was very computationally intensive. As an alternative, the DeeperCut model introduced by Insafutdinov *et al.* [32] appeared, which uses a much more robust detector and a better incremental optimisation strategy, resulting in better performance and computation time.

Later, multitasking structures were also used in this type of HPE. Papandreou *et al.* introduced PersonLab [33], which combined both person estimation and segmentation modules for keypoint detection and assignment. It consisted of three offsets: short range to refine the heatmaps; medium range to predict the different keypoints; and long range to group the keypoints into instances. A second example presented by Kocabas *et al.* consisted of a multi-task learning model with a residual network of poses, the MultiPoseNet [34], which was able to perform prediction, person recognition and semantic segmentation tasks simultaneously.

3.3 3D Single Person Pose Estimation

When we then come to the part that relates to three-dimensional estimation, this first sub-chapter begins by presenting a variety of techniques and examples of 3D SPPE, characterised by the fact that they may or may not be based on a particular model. At first glance, this definition may not be the most informative, but the following points attempt to provide a better context. It is equally important to add that this type of 3D estimation, like 2D SPPE, performs the recognition of the person in a first step, followed by their proper segmentation to identify their boundary in the respective image.

3.3.1 Model-Based

The parametric model is a powerful tool that provides important preliminary information that helps in the three-dimensional estimation of the human body, leading many approaches to resort to this alternative in the final reconstruction of a subject. In this context, and with recent technological advances, it is possible to divide this estimation into two different types. The first, the more traditional method based on optimisation, consists of fitting the parametric model of the human body to joint points or silhouettes, minimising a cost function. In the second case, with the current and successful application of Deep Learning techniques, researchers choose to estimate the human body from images through regression mechanisms [35].

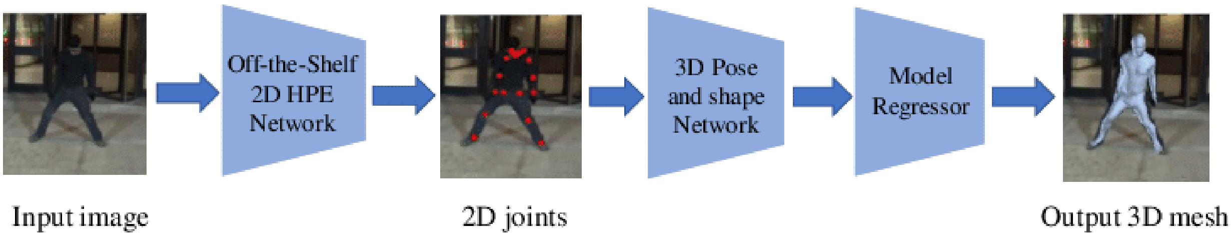


Figure 3.1: Estimation method based on a parametric model. Image taken from [14].

Optimization-based methods

The first methods based on the parametric model and optimisation focused on problems such as motion tracking. However, it was not long before work emerged that focused on estimating the morphology of the human body, based on fitting the parametric body to the silhouettes of single or multiple view images. However, they always proved to be less robust for the simple reason that they only worked in cases where the person was in a fixed position or that the models were still too simple to be able to reproduce the complicated deformation of the person's poses. It was only after the publication of the model SCAPE, already presented in the previous chapter, which allowed a more realistic representation of the human body. Balan *et al.* [36] proposed to use this tool to track the movement of a person in a three-dimensional space. In a first part, they first extracted the skeleton of the silhouettes and initialised SCAPE with the same skeleton. By re-fitting this previously initialised model to the silhouettes extracted from the images, they were able to deal with quite complex poses and obtain a much more robust three-dimensional model. Guan *et al.* [37] chose to construct a specific energy function using silhouettes and labelled 2D joint points. This allowed the SCAPE model to be adapted to these suggestions and, in turn, the created function to be minimised. Later, the advent of depth cameras, such as the Kinect, allowed the addition of depth information to strengthen the construction of the energy function, keeping in mind an improvement in the adaptation of the model. On the other hand, Bogo *et al.* [38] proposed the Delta model, an improved version of SCAPE, and used optimisation processes to adjust the joints and silhouettes to refine the appearance and also remove the inherent shifts in the depth images.

But the real change came with the publication of the article by A. Krizhevsky *et al.* [12]. From then on, Deep Learning techniques began to play a key role in many computer vision tasks, reaching levels of performance never seen before. One of his first investigations was an automatic method for estimating human pose and shape. The 2D joints were estimated by Deepcut [39] and then the SMPL model was fitted to these points to give a three-dimensional representation of the human body. It proved to be a very popular work, mainly because of its flexibility, as adding other factors such as silhouettes, 3D joint points from depth images and multiview images improved the results of previous methods. Alldieck *et al.* [40] presented a new solution to improve the estimation of a person's shape. By building an energy function between 16 multi-view silhouettes and an undesigned SMPL model, they were able to estimate some simple clothing items on the person's body. All this was possible because the non-designed model relaxed some of the constraints imposed by the original SMPL model. As a final example, Xu *et al.* [41] used deep neural networks to automatically determine the joint points and the corresponding 2D and 3D silhouettes, which made it possible to extract the skeleton of the pre-digitised model and fit it to the estimated points and silhouettes.

All this leads to the conclusion that optimisation-based methods depend heavily on a parametric model of the human body, used strictly to fit the, albeit limited, information extracted from images or videos. This is a classic approach that has proven useful in the reconstruction of the human body over the past decades. While the methods presented have been successful, they are slow and depend heavily on the accuracy of the information collected, which severely limits their application.

Regression-based methods

With the advent of Deep Learning and subsequent advances in estimating the pose of most public datasets, most of the published methods began to be based on DNN. Its excellent performance in tasks such as human pose segmentation and estimation enabled it to provide prior information such as joint points and silhouettes, and to predict prior suggestions in optimisation-based methods. In the context of the whole picture, however, the scenario became a little different, as these references became quite sparse and superficial. To solve the problem, the researchers began to use DNN to directly regress the pose and shape parameters of the entire image. In the early stages, factors such as silhouettes were often used to define the CNN loss function, which focuses on learning the shape of the person. For example, Dibra *et al.* [42] created a series of CNNs to regress silhouettes and learn the person's morphology based on the already known SCAPE model. The result was an improved model based on silhouettes at different scales and multiple views that eventually matched the shape of the originally proposed model. However, these two examples only deal with the human body in a simple pose, commonly known as A pose (see Figure 3.2) and are therefore not suitable for more complicated cases.

With the intention of overcoming this obstacle and thereby increasing flexibility in terms of pose, R. Cipolla *et al.* [43] proposed to use the SMPL model and derive its parameters from the loss resulting from the reprojection of the silhouettes. The success was so great that since then most methods have been using the SMPL model. Lassner *et al.* [44] created a loss function for the regression of the SMPL pose and shape parameters using about 91 2D joints obtained by SMPLify [45]. On the other hand, Kanazawa *et al.* [46] proposed an end-to-end restoration method of the human body. Using a CNN, they were also able to regress the SMPL parameters with only a single image. Basically, the training loss function was based on the joints of the respective image and the joints of the regressed SMPL model. This method achieved such good results that it inspired some proposed methods in the following years, namely [47] and [48]. In the first case, Pavlakos *et al.* integrated the silhouettes and a predicted mesh to improve shape performance in the construction of the respective loss function. Kolotouros *et al.* on the other hand, chose to combine the two methods for estimating the SMPL parameters, optimisation and regression, into a new method called SPIN, while incorporating SMPLify into a training loop to form a self-supervised structure.

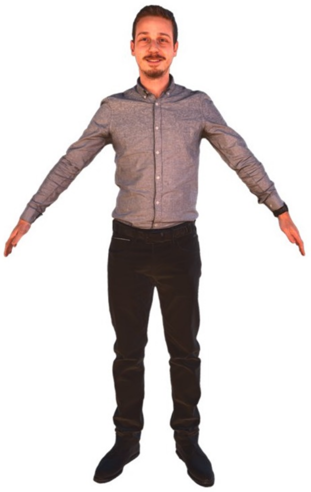


Figure 3.2: A Pose.

Despite all the incredible progress in the field of three-dimensional estimation, there is still one small drawback. The model-based methods described here, whether optimisation or regression, are hardly able to represent certain details of a person's appearance, since most parametric models of the human body do not take into account details such as clothing or hair.

3.3.2 Model-Free

Unlike previous 3D SPPE methods, the techniques presented below do not use parametric models of the human body as intermediate patterns or for its final three-dimensional representation. However, it is equally possible to divide them into two different classes: direct estimation approaches and 2D-to-3D lifting approaches.

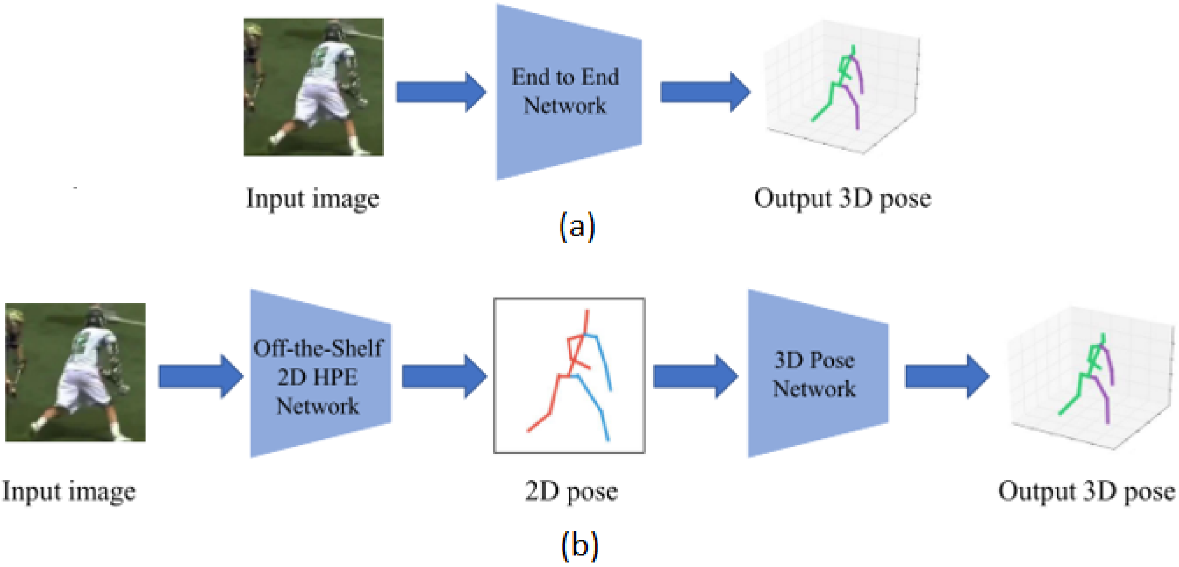


Figure 3.3: (a) Direct estimation approach. (b) 2D-to-3D lifting approach. Image taken from [14].

Direct estimation methods

Direct estimation methods, as shown in Figure 3.3.(a), infer the 3D pose of a human from simple images without estimating the 2D position representation in between. One of these early approaches used Deep Learning and was proposed by Li and Chan [49]. They used a shallow neural network that synchronously used small sliding windows to train the recognition of different body parts while regressing the coordinates of the person's pose. A similar system was presented by Li *et al.* [50] in which pairs of images and their respective three-dimensional poses served as input data for their network, so that the correct pairs could be assigned a high score and the rest a low score. However, the system was extremely inefficient as it required multiple inferences from the network. Sun *et al.* [51] have therefore opted for a structure-aware regression mechanism, i.e. instead of a representation based on joint points, they have chosen a representation based on bones, which has led to greater stability. To transform the non-linear 3D coordinate regression problem into a more manageable discrete form, Pavlakos *et al.* [52] proposed a volumetric representation in which the probability of voxels for each joint is predicted by a neural network.

2D-to-3D lifting methods

In the previous subchapter, the details of the 2D intermediate estimation of the human pose were discussed, albeit very briefly. This is also the case with the approaches presented below. Motivated by the recent success of 2D estimation, these 3D HPE methods use 2D pose detectors in an intermediate phase and then perform a 3D survey to determine the three-dimensional pose of the person. The first example proposed by Chen and Ramanan [53] was based on a set of DNNs responsible for the nearest neighbour correspondence between the predicted 2D pose and the corresponding 3D pose stored in a library, which could fail if the latter was not conditionally independent of the image. Martinez *et al.* [54] implemented a simple but highly effective residual network to regress the 3D joint positions considering the predicted 2D positions. On the other hand, instead of using 2D poses as an intermediate representation of their method, Zhou *et al.* [55] opted for HEMlets. The HEMlets, or Part-Centric Heatmap Triplets, used three heatmaps to represent the relative depth information of the joints of the extremities in relation to each part of the skeleton. This shortened the gap between 2D observation and 3D interpretation. Moreno-Noguer [56] captured the human pose by regressing two matrices. Both the distances of the 2D and 3D joints of the body were encoded in two Euclidean distance matrices, as these have the particularity of being invariant not only to rotations and translations of the image in the plane, but also to scale when normalisation operations are applied. In the case of Sharma *et al.* [57] and Li and Lee [58], they chose to create several hypotheses for three-dimensional poses and use a classification network to select the most useful posture.

However, the best known and most commonly used datasets in 3D HPE are usually developed in controlled environments. This makes the task of obtaining annotations on in-the-wild data quite difficult, not only because of the lack of data of this kind, but also because of the presence of unusual poses and occlusions. Nevertheless, this has not prevented the emergence of new work whose main focus has been on estimating 3D pose for data in-the-wild. As an example, Habibie *et al.* [59] adapted a projection loss to refine the human 3D pose without any kind of annotation. The developed 3D-2D projection module was responsible for computing the positions of 2D body joints when the three-dimensional pose was estimated in a previous layer of the network. Thus, the loss of the projection served to update the human 3D pose without the need for the usual annotations.

3.3.3 Conclusion

As for the repeatedly mentioned problem of partial occlusion of the person's limbs, it has already been noted that this is a very difficult issue in 3D HPE. The natural solution to overcome this problem was to estimate the human pose from multiple views, as the parts hidden in a certain angle may become visible in another angle. But in order to reconstruct the three-dimensional pose using this strategy, the corresponding location mapping between different cameras must be resolved. Furthermore, another relatively recent research field with limited work presented so far is the 3D Human Pose Estimation of multiple individuals. I have chosen not to go into too much detail on the topic as it strays a little from the aim of this dissertation, but they are essentially based on methods for estimating the 3D pose of a single person and Deep Learning architectures.

Table 3.1: Comparison of the 3D SPPE methods mentioned here in terms of used datasets, models and performance results.

Method	Year	Dataset	Approach	Type	MPJPE
[48]	2019	Human3.6M, MPI-INF-3DHP	Model based	Regression	41.1 mm
[59]	2019	Human3.6M, MPI-INF-3DHP, MPII	Model free	Lifting	49.2 mm
[55]	2019	Human3.6M, MPI-INF-3DHP	Model free	Lifting	39.9 mm
[58]	2019	MPI-INF-3DHP, MPII	Model free	Lifting	52.7 mm
[57]	2019	Human3.6M, HumanEva	Model free	Lifting	58.0 mm
[41]	2018	Human3.6M	Model based	Optimization	90.5 mm
[47]	2018	Human3.6M, UP-3D	Model based	Regression	75.9 mm
[46]	2018	COCO, MPII, LSP, Human3.6M	Model based	Regression	87.97 mm
[44]	2017	UP-3D, HumanEva, Human3.6M	Model based	Regression	80.7 mm
[43]	2017	UP-3D	Model based	Regression	-
[54]	2017	Human3.6M, HumanEva, MPII	Model free	Lifting	62.9 mm
[52]	2017	Human3.6M, HumanEva, MPII	Model free	Direct	71.9 mm
[53]	2017	Human3.6M	Model free	Lifting	82.7 mm
[51]	2017	Human3.6M, MPII	Model free	Direct	48.3 mm
[42]	2016	CAESAR	Model based	Regression	-
[56]	2016	Human3.6M, HumanEva-I, LSP	Model free	Lifting	87.3 mm
[50]	2015	Human3.6M	Model free	Direct	120.2 mm
[49]	2014	Human3.6M	Model free	Direct	132.2 mm

3.4 Telerehabilitation

Since telerehabilitation is the main topic of this dissertation, the next section presents a small selection of platforms of this kind. However, it is equally important to emphasise that these types of systems are quite new and are essentially based on motion detection mechanisms.

As mentioned earlier, physiotherapy is characterised by being a particularly expensive service that is laborious and involves long waiting times for a particular treatment. An example of this is the detection and correction of gait defects. However, gait rehabilitation can be supported and accelerated by regular exercises at home with the help of an automatic feedback solution. The first proposal by Ropars *et al.* [60] analysed the range of motion of participants' shoulders using motion capture data. The method assessed the hypermobility of the joints[1] of each shoulder, the main risk factor for their instability. At the same time, Bednarski and Bielak [61] used the same technology, but in the diagnosis of knee injuries. However, it also became possible to develop an alternative diagnostic system using HPE instead of motion capture. Kleanthous *et al.* [62] have thus shown that, in addition to diagnosis and rehabilitation, gait analysis also allows possible identification of imbalances and prediction of falls, facilitating a rapid response to these types of incidents.

Subsequently, gait analysis proved to be equally effective in quantitatively comparing the performance of athletes, especially in sports such as athletics. As with the rehabilitation exercises offered by the physiotherapist, correct posture and technique in training is crucial to achieve the highest standards and avoid injury and future health problems. In this way, the integration of HPE techniques into these types of platforms offers doctors and sports coaches the opportunity to analyse the biomechanics of their patients or athletes and help them become more effective in their treatment or sport.

[1] Hypermobility syndrome represents a set of clinical symptoms in which the patient is able to perform larger than normal movements in the joints, mainly in the hands, shoulders and knees.

4 Proposed Approach

After contextualising some of the methods available in the literature, this next chapter delves into the methods used in the practical part of this thesis. Thus, an initial analysis of the human body model used is presented, followed by a detailed description of the estimation approach used, which includes all the considerations for its development, the theoretical foundations, the mathematical calculations applied and the expected benefits.

4.1 Skinned Multi-Person Linear Model

As mentioned above, the present work focuses on the three-dimensional estimation of a patient's posture and body shape in order to study and segment the curvature of the back when performing telerehabilitation exercises. Therefore, a very well known parametric model called Skinned Multi-Person Linear is used. The main objective of the system presented by Loper et al. [19] is to create a realistic and animated human body that can be naturally deformed according to the pose assumed by the person, while at the same time representing the movements of certain soft tissues such as muscles and synovial tissues[1] of a real person. It is equally important to add that all this is possible on CPU with standard rendering engines, and at a much higher speed.

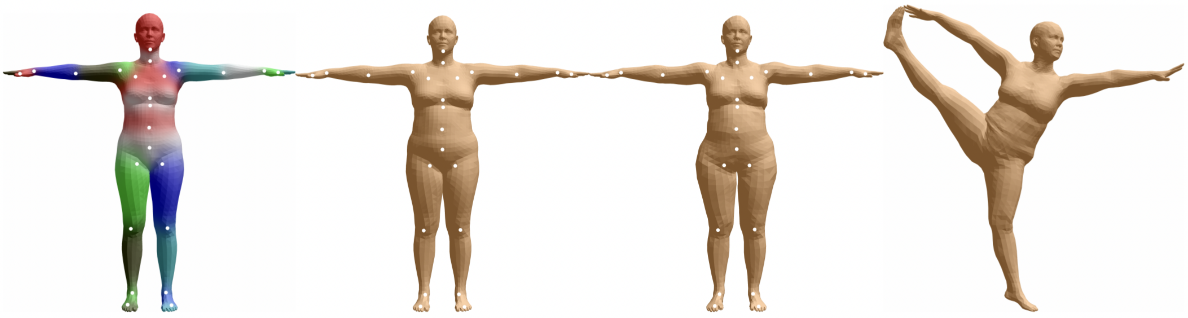
With a more technical approach, SMPL consists of a vertex-based blend skinning process parameterised by shape, β, and pose, θ, parameters. The first, relating to shape, is interpreted as a set of values that respond to an increase or decrease in certain initial values. The next parameter, pose, is determined by the relative rotations of all parts of the body, the latter being defined by the joints available in a kinematic tree [19]. The term kinematic tree thus means a skeletal structure of limbs connected by such joints, each of which is influenced by the preceding limb, commonly called the father, and eventually influences the next, the child.

[1] Lining that surrounds the joints and forms a joint capsule. The cells of the synovial tissue produce a small amount of fluid that nourishes the cartilage and reduces friction, which facilitates movement.

The blend skinning process is known to add a mesh surface, in this case a triangular one, to the underlying skeletal structure, transforming each of its vertices by the influence of the adjacent bones. To follow standard skinning practise, the model is defined by an average model shape represented by 23 joints, a concatenated vector of $N = 6890$ vertices $\bar{T} \in R^{3n}$ in an initial position, $\vec{\theta}^*$, and a set of weights, $\mathcal{W} \in R^{N \times K}$, shown in Figure 4.1(a). In the case of figure 4.1(b), the function $B_S(\vec{\beta})$ is introduced which receives a vector of shape parameters, $\vec{\beta}$, and returns the distance of the newly obtained shape compared to the previous standard model. For figure 4.1(c), a third function $B_P(\vec{\theta})$ is added which receives a vector of pose parameters, θ, and takes into account the deformation effects caused by them. The pose vector is given by $\vec{\theta} = [\vec{w}_0^T, ..., \vec{w}_K^T]^T$, where $\vec{w}_k \in R^3$ corresponds to the axis-angle representation of the relative rotation of part k with respect to its parent in the kinematic tree, and the angle relative to the axis of each joint j being converted into a rotation matrix by the Rodriguez Formula (Equation 4.1). Finally, in 4.1(d), the blend skinning function $W(\bar{T}, J, \vec{\theta}, \mathcal{W})$ is applied, which assumes the vertices in the initial position $\bar{T}$, the position of each of the joints J, the pose $\vec{\theta}$ and a series of weights $\mathcal{W}$ to rotate the vertices about the estimated joints [19].

$$exp(\vec{w}_j) = \mathcal{I} + \hat{w}_j \sin(\|\vec{w}_j\|) + \hat{w}_j^2 \cos(\|\vec{w}_j\|), \tag{4.1}$$

where $\mathcal{I}$ is the 3×3 identity matrix.



(a) $\bar{\mathbf{T}}, \mathcal{W}$ (b) $\bar{\mathbf{T}} + B_S(\vec{\beta}), J(\vec{\beta})$ (c) $T_P(\vec{\beta},\vec{\theta}) = \bar{\mathbf{T}}+B_S(\vec{\beta})+B_P(\vec{\theta})$ (d) $W(T_P(\vec{\beta},\vec{\theta}), J(\vec{\beta}), \vec{\theta}, \mathcal{W})$

Figure 4.1: Example of human body pose and shape configurations calculated by the SMPL model. (a) Base model without any input parameters, with blend weights illustrated in colour, and joints in white. (b), (c) Parametrized models by introducing some parameters. In figure (c), an expansion of the hips can be observed, for instance. (d) Final model with its respective deformed vertices.

The SMPL model was created from 3D scans of different faces and poses of thousands of people, both independent of each other, using Principal Component Analysis (PCA). By using PCA, we can regress the different scans back to the ten main orthonormal shape components $\mathcal{S}$ and 72 pose displacements $\mathcal{P}$. For example, the first and second principal components of the parameters of $\mathcal{S}$ relate to changes in size and weight respectively. With the matrices $\mathcal{S}$ and $\mathcal{P}$ we can define the functions B_S, equation 4.2, and B_P, equation 4.3. The first helps to create a more realistic mesh, while the second prepares the subject for a pose.

$$B_S(\vec{\beta};\mathcal{S}) = \sum_{n=1}^{|\vec{\beta}|} \beta_n S_n \tag{4.2}$$

$$B_P(\vec{\theta};\mathcal{P}) = \sum_{n=1}^{|9\vec{K}|} (R_n(\vec{\theta}) - R_n(\vec{\theta}^*))P_n \tag{4.3}$$

Where $S_n, P_n \in R^{3N}$, and $R_n(\vec{\theta})$ is the 3×3 Rodrigues matrix.

Finally, since the SMPL model is differentiable and thus can be used with deep networks [63], it serves as an illustration tool for the applied neural network, whose task is to find the correct shape and coefficients of an image or video and thus create an accurate human representation.

4.2 Hybrid Analytical-Neural Inverse Kinematics Solution

Apart from this, the reconstruction of a 3D surface from monocular images is one of the oldest problems in computer vision. Since the publication of statistical parametric models of human body shape such as SMPL, the recovery of the three-dimensional mesh of individuals has attracted more and more attention. With the aim of obtaining increasingly well-matched and physically plausible results, the two aforementioned paradigms have been developed. Optimisation-based approaches, where different data and regularisation conditions are explored as optimisation targets, and regression-based approaches, where deep learning techniques are used. However, in the first case, not only are the results sensitive to initialisation, but the optimisation problem is also non- convex and requires too much time to solve. On the other hand, in regression methods, since the parameter space of the statistical model is abstract, it is difficult for networks to learn the mapping function.

These challenges have led a group of researchers to focus on the area of three-dimensional key point estimation, which relates to methods that precede direct regression and whose high performance comes from adopting volumetric heat maps as the final representation when learning the positions of 3D joints. This inspired them to associate 3D joints with the parametric model of the human body, Figure 4.2. Not only because joints facilitate the estimation of the volumetric mesh, but also because current methods for estimating key points lack an explicit modeling of the length distribution of the body's bones, which results in an unrealistic prediction of body structures, such as abnormal limb proportions. Thus, by leveraging the parametric body model, the represented human form fits better to the real human body.

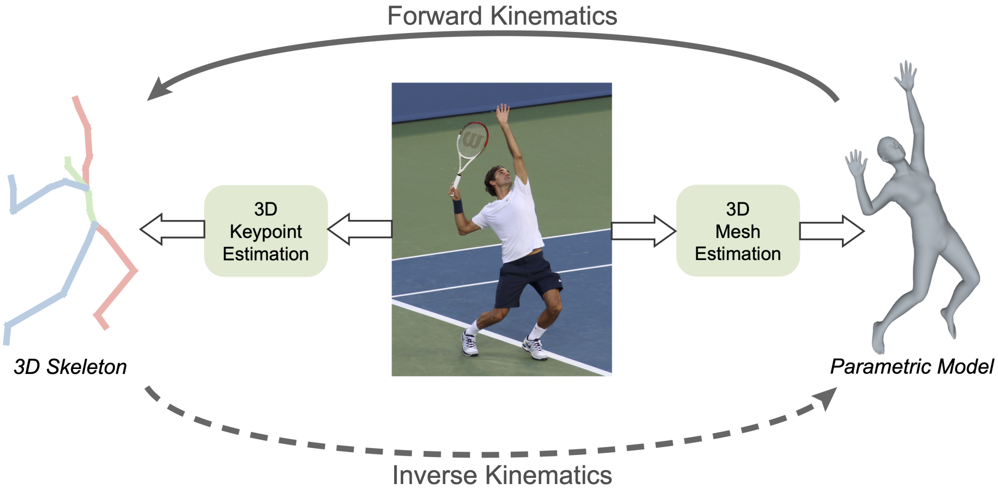


Figure 4.2: Loop between 3D skeleton and parametric model, via HybrIK. The 3D skeleton projected by the neural network can be transformed into a body mesh by inverse kinematics without sacrificing precision. In turn, the parametric body mesh can generate a realistic 3D structural skeleton through direct kinematics [64].

Thus was born HybrIK [64], a hybrid analytical-neural inverse kinematics solution that bridges the gap between the two previous types of estimation. More specifically, inverse kinematics is a mathematical process responsible for finding the relative rotations that allow the desired positions of the body's joints to be generated, and is considered a problem without a single solution. In this way, the core of the approach is to propose an innovative IK solution through a twist-and-swing decomposition. That is, each part of the skeleton is decomposed into a longitudinal rotation and a plane rotation. In Hybrik, these rotations are recursively composed along the kinematic tree, where the swing rotation is calculated analytically and the twist rotation is predicted. A key feature of this method is that the relative rotation estimated by HybrIK is naturally aligned with the 3D skeleton, without the need for additional optimisation procedures as in previous approaches. Furthermore, all of its operations are differentiable, which enables the simultaneous training of 3D joints and volumetric meshes of the human body.

4.2.1 Preliminary

Before I present the final architecture, it is important to explain some terms and mathematical calculations behind the HybrIK model.

Forward Kinematics

Forward kinematics (FK) in the context of human posture generally refers to the process of calculating the reconstructed pose $Q = \{q_k\}_{k=1}^{K}$ with the resting pose model $T = \{t_k\}_{k=1}^{K}$ and the set of relative rotations $R = \{R_{pa(k),k}\}_{k=1}^{K}$ as input:

$$Q = FK(R, T), \tag{4.4}$$

where K corresponds to the number of joints, $q_k \in \mathbb{R}^3$ denotes the reconstructed 3D position of the k-th joint, $t_k \in \mathbb{R}^3$ denotes the position of the k-th joint of the residual pose model, $pa(k)$ returns the index of the parent joint of the k-th joint and $R_{pa(k),k}$ corresponds to the relative rotation of the k-th joint relative to its parent joint. The FK can be performed by recursively rotating the model body part from the root joint to the leaf joints:

$$q_k = R_k(t_k - t_{pa(k)}) + q_{pa(k)}, \tag{4.5}$$

where $R_k \in \mathbb{SO}(3)$ is the global rotation of the k-th joint relative to the canonical space of the rest pose. The global rotation can be calculated recursively:

$$R_k = R_{pa(k)} R_{pa(k),k}. \tag{4.6}$$

For the root joint, which has no relatives, we have $q_0 = t_0$.

Inverse Kinematics

Inverse kinematics (IK) is the reverse process of FK and calculates the relative rotations R that can produce the desired joint positions of the input body $P = \{p_k\}_{k=1}^{K}$. This process can then be formulated as follows:

$$R = IK(P, T), \tag{4.7}$$

where p_k denotes the k-th joint of the entry position. Ideally, the resulting rotations should fulfil the following condition:

$$p_k - p_{pa(k)} = R_k(t_k - t_{pa(k)}) \qquad \forall 1 \leq k \leq K, \tag{4.8}$$

Similarly, we have $p_0 = t_0$ for the root joint that has no parent. But the IK problem is ill-posed because there is no concrete solution or, on the contrary, there are too many solutions that fit the locations of the target joints.

Twist-and-swing Decomposition

Usually, in the analytical IK formulation, some joints of the body are given fewer degrees of freedom (DoFs) to simplify the problem, for example 1 or 2 DoFs. In the case of HybrIK, it was assumed that each joint of the body has 3 full degrees of freedom. As shown in Figure 4.3, a $R \in \mathbb{SO}(3)$ rotation can be decomposed into a twist rotation, R^{tw}, and a swing rotation, R^{sw}. Given the initial body part vector of the model and the target vector, the process of solving R can be formulated as follows:

$$R = \mathcal{D}(\vec{p}, \vec{t}, \phi) = \mathcal{D}^{sw}(\vec{p}, \vec{t})\mathcal{D}^{tw}(\vec{t}, \vec{\phi}) = R^{sw}R^{tw}, \tag{4.9}$$

where ϕ is the angle of twist estimated by a neural network, $\mathcal{D}^{sw}(\cdot)$ is a closed-form solution of the swing rotation, and $D^{tw}(\cdot)$ transforms ϕ into twist rotation. Here R must fulfil the condition of the Equation 4.8, i.e., $\vec{p} = R\vec{t}$.

- **Swing:** the swing rotation has the $\vec{n}$ axis perpendicular to $\vec{t}$ and $\vec{p}$. Therefore, it can be formulated as follows:

$$\vec{n} = \frac{\vec{t} \times \vec{p}}{\|\vec{t} \times \vec{p}\|} \tag{4.10}$$

and the swing angle α satisfies:

$$\cos\alpha = \frac{\vec{t} \cdot \vec{p}}{\|\vec{t}\|\|\vec{p}\|}, \qquad \sin\alpha = \frac{\|\vec{t} \times \vec{p}\|}{\|\vec{t}\|\|\vec{p}\|}. \tag{4.11}$$

Therefore, the closed-form solution of the swing rotation R^{sw} can be derived by Rodrigues' formula:

$$R^{sw} = \mathcal{D}^{sw}(\vec{p}, \vec{t}) = \mathcal{I} + \sin\alpha[\vec{n}]_\times + (1 - \cos\alpha)[\vec{n}]^2_\times, \tag{4.12}$$

where $[\vec{n}]_\times$ is the asymmetric matrix of $\vec{n}$ and $\mathcal{I}$ is the 3×3 identity matrix.

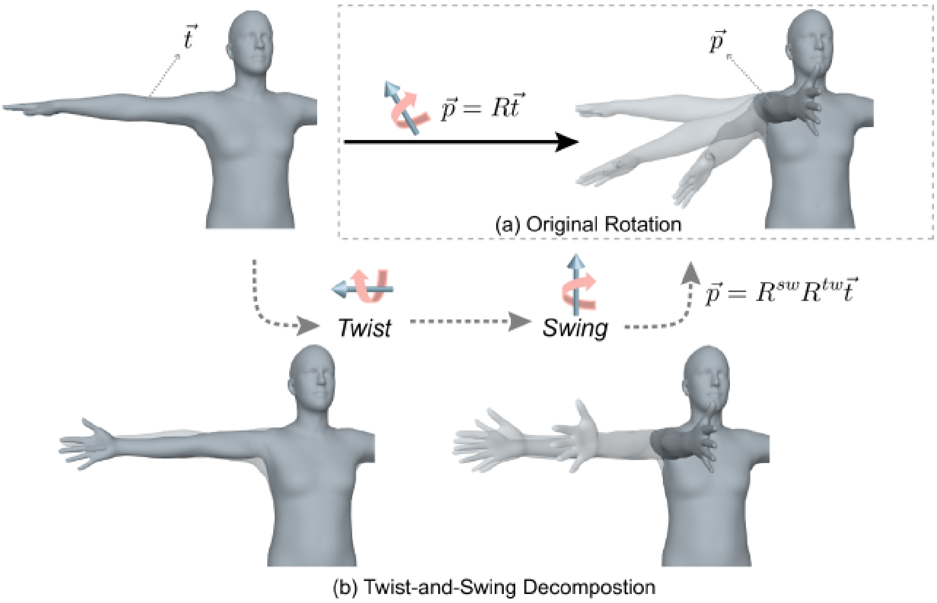


Figure 4.3: Illustration of the twist-and-swing decomposition. (a) In the original rotation, the right palm is turned downwards, forwards and to the left in a single step. (b) In the twist-and-swing decomposition, the rotation is divided into two steps: First the palm is turned 90° and then the whole hand is moved forward [64].

- **Twist:** the twist rotation revolves around itself. So with $\vec{t}$ itself as the axis and ϕ as the angle, we can determine the rotation of the twist R^{tw}:

$$R^{tw} = \mathcal{D}^{tw}(\vec{t}, \phi) = \mathcal{I} + \frac{\sin\phi}{\|\vec{t}\|}[\vec{t}]_\times + \frac{(1-\cos\phi)}{\|\vec{t}\|^2}[\vec{t}]_\times^2, \tag{4.13}$$

where $[\vec{t}]_\times$ is the asymmetric matrix of $\vec{t}$.

Since the functions $\mathcal{D}^{sw}$ and $\mathcal{D}^{tw}$ are fully differentiable, it is possible to integrate the twist-and-swing decomposition into the training process. Although a neural network is required to learn the twist angle, the difficulty of learning is greatly reduced. In addition, due to the physical limitations of the human body, the angle of twist has a small range of variation. Therefore, it is also easier to learn the mapping function.

4.2.2 Naive HybrIK

The IK process, like the FK process, can run recursively through the kinematic tree. The first step is to determine the rotation of the global root R_0, whose closed-form solution uses Singular Value Decomposition (SVD), and the positions of the spine and the left and right hips. Then, at each step k, it is assumed that the rotation of the parent joint $R_{pa(k)}$ is known. We can therefore reformulate Equation 4.8 with Equation 4.6 as follows:

$$R_{pa(k)}^{-1}(p_k - p_{pa(k)}) = R_{pa(k),k}(t_k - t_{pa(k)}). \tag{4.14}$$

Whether $\vec{p}_k = R_{pa(k)}^{-1}(p_k - p_{pa(k)})$ and $\vec{t}_k = (t_k - t_{pa(k)})$, it is possible to solve the relative rotation by Equation 4.9:

$$R_{pa(k),k} = \mathcal{D}(\vec{p}_k, \vec{t}_k, \phi_k), \tag{4.15}$$

where ϕ_k is the twist angle predicted by the network for the k-th joint. The set of twist angles is denoted $\Phi = \{\phi_k\}_{k=1}^{K}$. Since the rotation matrices are orthogonal, their inverse is equal to their transpose, i.e. $R_{pa(k)}^{-1}(p_k - p_{pa(k)}) = R_{pa(k)}^{T}(p_k - p_{pa(k)})$, which keeps the resolution process differentiable.

This procedure was called Naive HybrIK, whereby it was possible to solve for the relative rotation $R_{pa(k),k}^{-1}$ instead of the global rotation R_k. The reason for this is quite simple: if the global rotation were decomposed directly, the resulting twist angle would depend on the rotations of all the predecessor rotations along the kinematic tree, which would lead to increased variation in the distal joints of the limbs and learning difficulties on the part of the network.

4.2.3 Adaptive HybrIK

Although the above procedure seems effective, it follows an unstated assumption: $\|p_k - p_{pa(k)}\| = \|t_k - t_{pa(k)}\|$. Otherwise, there would be no solution to Equation 4.8. Unfortunately, in this case, the body parts predicted by the 3D keypoint estimation method do not always agree with the resting pose model. In Naive HybrIK, Equation 4.9 can still be solved because the condition is transformed into:

$$p_k - p_{pa(k)} = R_k(t_k - t_{pa(k)}) + \vec{\epsilon}_k, \tag{4.16}$$

where $\vec{\epsilon}_k$ denotes the error in the k-th step, which has the same direction as $p_k - p_{pa(k)}$ and $\|\vec{\epsilon}_k\| = |\|p_k - p_{pa(k)}\| - \|t_k - t_{pa(k)}\||$. To analyse the reconstruction error, the difference between the input pose P and the reconstructed pose Q was compared:

$$\|P - Q\| \Leftrightarrow \sum_{k=1}^{K} \|p_k - q_k\|, \tag{4.17}$$

where $Q = FK(R,T) = FK(IK(P,T),T)$. Combining Equation 4.6 and Equation 4.16 we get:

$$\begin{aligned} p_k - q_k &= p_{pa(k)} - q_{pa(k)} + \vec{\epsilon}_k \\ &= p_{pa^2(k)} - q_{pa^2(k)} + \vec{\epsilon}_{pa(k)} + \vec{\epsilon}_k \\ &= ... = \sum_{i \in A(k)} \vec{\epsilon}_i, \end{aligned} \tag{4.18}$$

where $pa^2(k)$ denotes the parent index of the pa(k)-th joint and $A(k)$ the set of ancestors of the k-th joint. This means that the difference between the input joint p_k and the reconstructed joint q_k is accumulated along the kinematic tree, leading to more uncertainty in the distal joint.

To solve the problem of error accumulation, a second method, Adaptive HybrIK, has been proposed. In this method, the target vector is adaptively updated by the newly reconstructed original joints. Let $\vec{p}_k = R^{-1}_{pa(k)}(p_k - q_{pa(k)})$ and $\vec{t}_k$ be equal to Naive HybrIK. In this way, the condition in Adaptive HybrIK can be formulated as follows:

$$p_k - q_{pa(k)} = R_k(t_k - t_{pa(k)}) + \vec{\epsilon}_k. \tag{4.19}$$

So we have:

$$\begin{aligned} p_k - q_{pa(k)} &= q_k - q_{pa(k)} + \vec{\epsilon}_k \\ &\Rightarrow p_k - q_k = \vec{\epsilon}_k. \end{aligned} \tag{4.20}$$

Compared to the previous solution, Equation 4.18, the reconstructed error of this new solution depends only on the current joint and is not accumulated from the previous joints. As can be seen in Figure 4.4, in Naive HybrIK the descending joints continue this error once the main joint is out of position. In the case of Adaptive HybrIK, the solution of relative rotation always points to the target joint, reducing the error. A solution would go through an iterative global optimisation process, but this is neither differentiable nor does it allow end-to-end training.

4.2.4 Framework

The structure of the followed approach is shown in Figure 4.5. First, a heat map is generated using the neural network through its deconvolution layers and used to predict the three-dimensional joints P. At the same time, the twist angle Φ and the shape parameters β are learned from the visual cues of the fully-connected layers. Secondly, the shape parameters are used to obtain the rest position T by the SMPL model. By combining P, T and Φ, it is then possible to run HybrIK to solve the relative rotations R of the 3D pose, i.e. the

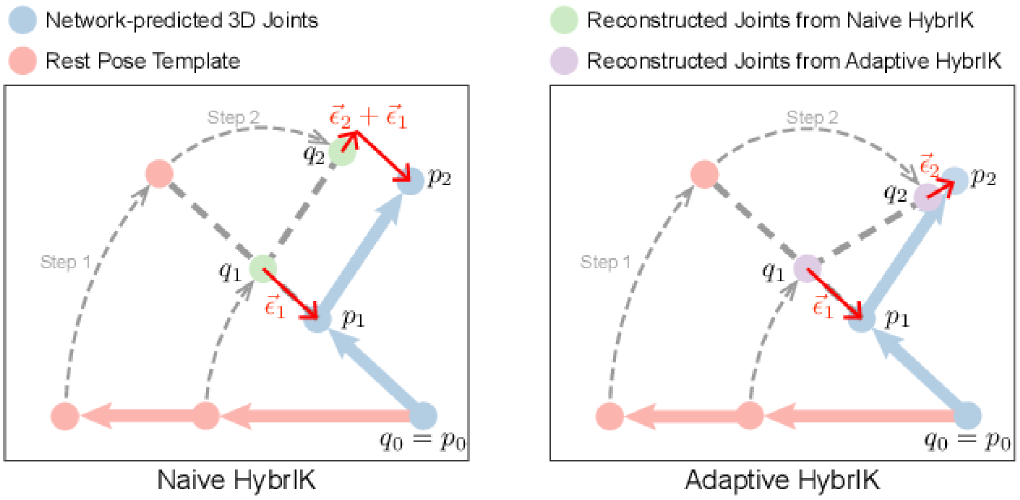

Network-predicted 3D Joints
Rest Pose Template
Reconstructed Joints from Naive HybrIK
Reconstructed Joints from Adaptive HybrIK
Step 1
Step 2
$\vec{\epsilon}_2 + \vec{\epsilon}_1$
$\vec{\epsilon}_1$
$\vec{\epsilon}_2$
q_1
q_2
p_1
p_2
$q_0 = p_0$
Naive HybrIK
Adaptive HybrIK

Figure 4.4: Example of the reconstruction error. The resting pose is rotated in two stages to q_1 and q_2. In the first step, the reconstruction error is $\vec{\epsilon}_1$ due to the inconsistency of the bone length. In the second step, Naive HybrIK takes $p_2 - p_1$ as the target direction, resulting in a cumulative error of $\vec{\epsilon}_1 + \vec{\epsilon}_2$. Instead, Adaptive HybrIK selects the reconstructed joint q_1 to form the target direction $p_2 - q_1$, reducing the error to only $\vec{\epsilon}_2$ [64].

position parameters θ. Finally, using the function $\mathcal{M}(\theta, \beta)$ of the SMPL model, it is possible to obtain the triangle mesh M.

Heatmaps and 3D Keypoint Estimation

The network used to estimate the heatmaps and the respective three-dimensional keypoints was the HRNet-W48 [13], initialised with pre-trained weights from ImageNet and adapted from the ResNet design by distributing the depth to each stage and the number of channels for each resolution. Its output is fed into a medium pooling layer, followed by the fully-connected layers to regress β and ϕ. The implementation was done in PyTorch and the input images were scaled down to 256×256. In addition, the learning rate is initially set to 1×10^{-3} and reduced by a factor of 10 in the 90th and 120th epochs, giving a total of 140 epochs. The optimisation function chosen was *Adam*.

Twist Angle Estimation

Instead of regressing the scalar value ϕ_k directly, the authors decided to learn a two-dimensional vector $(cos\phi_k, sin\phi_k)$ to avoid the problem of discontinuity. The $l2$ loss is applied:

$$\mathcal{L}_{tw} = \frac{1}{K}\sum_{k=1}^{K} \|(cos\phi_k, sin\phi_k) - (cos\hat{\phi}_k, sin\hat{\phi}_k)\|^2, \tag{4.21}$$

where $\hat{\phi}_k$ is the actual angle of twist for the k-th joint.

Collaboration with SMPL

The SMPL model makes it possible to obtain the skeleton resting pose with the additive displacements according to the shape parameters β:

$$T = W(\bar{M}_T + B_S(\beta)), \tag{4.22}$$

where $\bar{M}_T$ are the mesh vertices of the mean resting pose and $B_S(\beta)$ is the shape blending function provided by SMPL. Then the pose parameters, θ, are calculated by HybrIK in a differentiable manner. In the training phase, the shape parameters β and the rotation parameters θ are supervised as follows:

$$\mathcal{L}_{shape} = \|\beta - \hat{\beta}\|^2, \qquad \mathcal{L}_{rot} = \|\theta - \hat{\theta}\|^2. \tag{4.23}$$

The overall learning loss is thus given by:

$$\mathcal{L} = \mathcal{L}_{pose} + \mu_1\mathcal{L}_{shape} + \mu_2\mathcal{L}_{rot} + \mu_3\mathcal{L}_{tw}, \tag{4.24}$$

where μ_1, μ_2 and μ_3 correspond to the weights of the loss items.

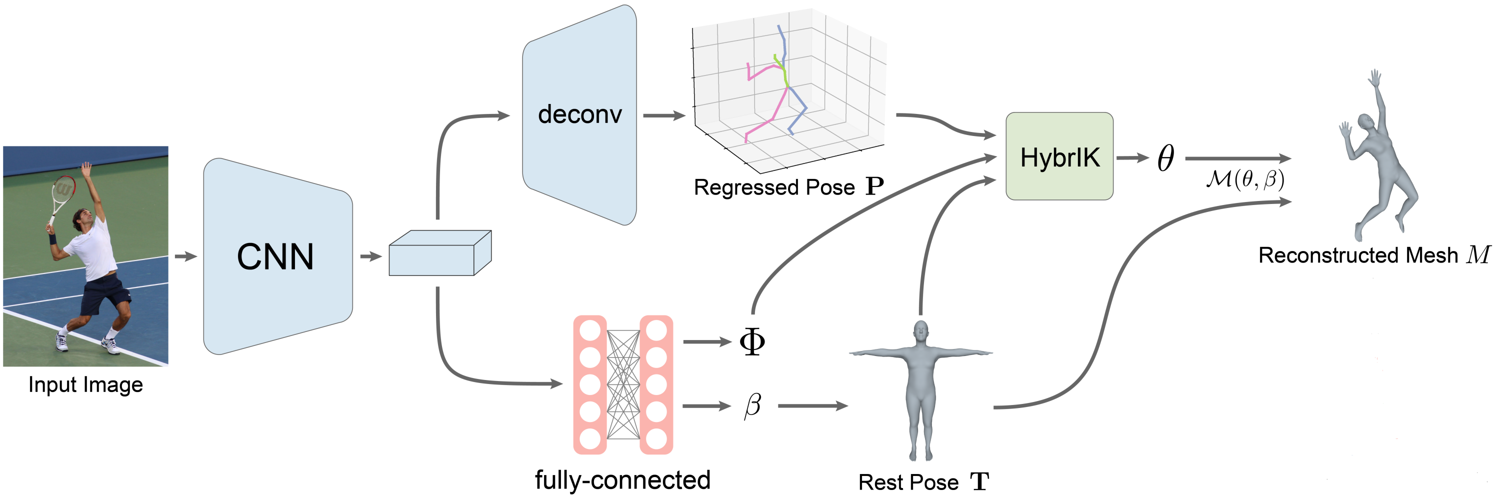


Figure 4.5: Overview of the used framework.

5 Results and Discussion

In this chapter, we describe the datasets used for training and quantitative evaluation of HybrIK. Then, some experiments are conducted to analyse the proposed concepts and some results of comparing the model with approaches already existing in the literature are reported. Finally, a second set of results derived from the application of the structure 4.5 to examples of rehabilitation exercises is presented, as well as a detailed explanation of the whole process behind the segmentation of the patient's spinal curve.

5.1 Datasets

The datasets used were Human3.6M [65], MPI-INF-3DHP [66], COCO [67] and 3D Poses in the Wild (3DPW) [68]. Each of these datasets contains annotated images of poses of people engaged in sports or activities of daily living.

The Human3.6M is a multiview dataset recorded in an enclosed and controlled space that serves as a reference for estimating the 3D human pose. The data is organised into videos of multiple people performing different activities such as walking, discussing or eating. The images were extracted from these videos using the tools provided by OpenCV [69]. In the specific case of our method, subjects S1, S5, S6, S7 and S8 were used for training and only S9 and S11 for testing.

Similarly, MPI-INF-3DHP also corresponds to a multiview dataset containing videos of six people performing seven actions indoors. The data is in the same video format and was extracted as in Human3.6M. However, in this case, both the training and testing sets were used in their entirety for training and testing HybrIK respectively. On the other hand, COCO is characterised by the fact that it is a large dataset, essentially used for object detection and segmentation, containing images and annotations in JSON format of people engaged in a variety of outdoor activities. It is only used for training.

The last, 3DPW, also outdoors, contains precise 3D poses for evaluation, i.e. it has 2D and 3D pose annotations and camera poses for each image sequence. As with Human3.6M, the data is also organised by videos in which a person is performing activities such as walking, discussing or performing a physical activity. In addition, it is able to provide highly accurate SMPL ground truth parameters and is only used for model evaluation.

5.2 Ablation Study

In order to assess the efficiency of the twist-and-swing decomposition and the HybrIK algorithm, a series of studies were conducted by the authors of HybrIK.

5.2.1 Analysis of the twist rotation

To demonstrate the effectiveness of the twist-and-swing decomposition, the first step was to analyse the distribution of the twist angle in the 3DPW test set, shown in Figure 5.1. As expected, due to the physical limitation, only the neck, elbows and wrists have a wide range of variation. All other joints have a limited range for the angle of twist of less than 30°, which can be estimated with relative confidence.

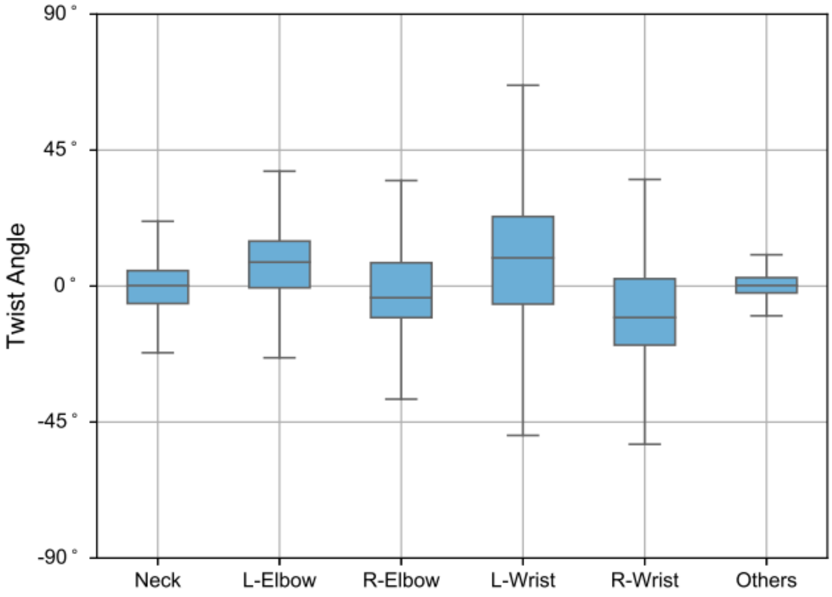


Figure 5.1: Distribution of the twist angle. Only a few joints have a range over 30º. Other joints have a limited range of twist angle.

The next step was to develop a test to find out how the twist angles would affect the reconstructed pose and shape. Thus, the 24 basic joints of the SMPL and the shape parameters served as input for HybrIK and with regard to the twist angle, the random values in $[-\pi, \pi]$ and the values estimated by the network were compared. Table 5.1 thus shows the results of the mean error of the 24 reconstructed SMPL joints, the 14 LSP joints, the body mesh and the angle of twist. Note that the LSP joints are obtained from the body mesh by a pre-trained regression mechanism, as for [46], [48]. A brief analysis shows that the predicted twist angles significantly reduce the error at the mesh vertices and at the LSP joints. Since most of the twist angles are close to zero, as presented in Figure 5.1), equating all twist angles to zero also leads to acceptable performance. It can also be observed that the incorrect angles of twist have no significant effect on the reconstructed SMPL joints (first value on the left side), but only the swing rotations.

	Random Twist			Estimated Twist			Zero Twist		
	24 jts	14 jts	Vert.	24 jts	14 jts	Vert.	24 jts	14 jts	Vert.
Error	0.1	40.0	67.3	0.1	6.1	10.0	0.1	6.8	12.1

Table 5.1: Reconstruction error of the 24 reconstructed SMPL joints, the 14 LSP joints, and the body mesh, with varying twist angles.

5.2.2 Robustness of HybrIK

To demonstrate the better performance of Adaptive HybrIK over Naive HybrIK, the error of the reconstructed joints of the two was compared. Thus, each of the IK algorithms first received the real joints, the angle of twist and the shape parameters as input to check whether they really introduced additional errors. Next, it was decided to assign noise to the same inputs. As can be seen in Table 5.2, both algorithms have negligible errors when the joints are correct. However, with noisy joints, Naive HybrIK accumulates a much higher error along the kinematic tree than Adaptive HybrIK.

	GT Joints	± 10 mm	± 20 mm	± 30 mm
Naive HybrIK	0.1	16.2	34.0	53.4
Adaptive HybrIK	0.1	9.8	20.2	31.2

Table 5.2: Naive vs. Adaptive with different input joints. MPJPE of 24 joints is reported.

5.2.3 Error correction capability of HybrIK

In this experiment, the ability of the HybrIK algorithm to correct errors was investigated. Thus, the algorithm is fed with the 3D joints, twist angles and shape parameters predicted by the high-resolution network. In addition, the SMPLify algorithm [45] was applied to the predicted pose to have a comparison term. As can be seen from the Table 5.3, the error of the reconstructed joints is reduced to 79.2 mm after HybrIK, while it increases to 114.3 mm with SMPLify. This is because the network can predict unrealistic body postures, such as left-right asymmetry or abnormal limb proportions. In contrast, because the resting pose is generated by the parametric body model, it is ensured that the reconstructed pose matches the realistic distribution of the body shape.

	Predicted Pose	HybrIK	SMPLify
MPJPE (24 jts) ↓	88.2 mm	79.2 mm	114.3 mm

Table 5.3: Error correction capability of HybrIK compared to the predicted pose and the SMPLify method.

5.3 Comparison with the State-of-the-art

To provide a fair comparison with previous methods for three-dimensional human pose and shape estimation, was used a pre-trained regressor that predicted the 14 LSP joints of the body mesh for evaluation in the 3DPW and Human3.6M datasets and 17 joints for the MPI-INF-3DHP.

Table 5.4 shows the results obtained, including model-based and model-free methods. According to the researchers behind HybrIK, it outperforms all previous methods on all three datasets and even improves the PVE of 21.9 mm on 3DPW. This only demonstrates that it is very accurate and reliable in restoring the body mesh using inverse kinematics.

5.3.1 Mean Per Joint Position Error (MPJPE) and Per Vertex Error (PVE)

Mean per joint position error (MPJPE) is the most common evaluation metric in 3D human pose estimation. Usually, the position error per joint is given by the Euclidean distance between the base joints and the predicted joints. The lower the value, the better the joint estimate. To calculate this error, the alignment of the root joints must be the same. In this thesis, the pelvic joint was chosen as the root, and all samples used are normalised and centred on the same root. The following equation defines our metric as follows:

$$MPJPE = \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N} \| J_{gt}^{(t)} - \hat{J}_{pred}^{(t)} \|^2, \tag{5.1}$$

where T is the number of samples, and N the number of joints.

It is also possible to use the same formula to evaluate all 6890 vertices of the body mesh predicted by SMPL. In this case, the actual SMPL parameters must be known in advance, and the only dataset available for this purpose is 3DPW. This metric is called Per Vertex Error (PVE).

5.3.2 Percentage of Correct Keypoints (PCK) and Area Under Curve (AUC)

In the PCK accuracy metric, the detected joint is considered correct if the distance between its predicted position and its actual position is less than a certain threshold. It is usually defined relative to the scale of the object within the bounding box. The area under the curve (AUC) is then calculated for a certain range of PCK thresholds.

These metrics are more meaningful and robust than the MPJPE and show more accurate prediction errors for each joint. The limit chosen was about 150 mm, which is about half the size of a person's head.

	3DPW			Human3.6M		MPI-INF-3DHP		
Method	PA-MPJPE	MPJPE	PVE	PA-MPJPE	MPJPE	PCK	AUC	MPJPE
SMPLify	-	-	-	82.3	-	-	-	-
HMR	81.3	130.0	-	56.8	88.0	72.9	36.5	124.2
Kolotouros et al.	70.2	-	-	50.1	-	-	-	-
Pavlakos et al.	-	-	-	75.9	-	-	-	-
Arnab et al.	72.2	-	-	54.3	77.8	-	-	-
SPIN	59.2	96.9	116.4	41.1	-	76.4	37.1	105.2
Moon et al.	58.6	93.2	-	41.7	55.7	-	-	-
Naive HybrIK	49.0	80.2	94.6	35.3	55.8	85.9	41.7	91.5
Adaptive HybrIK	**48.8**	**80.0**	**94.5**	**34.5**	**54.4**	**86.2**	**42.2**	**91.0**

Table 5.4: Benchmark of state-of-the-art models on 3DPW, Human3.6M and MPI-INF-3DHP datasets.

5.4 Spine curvature interpolation

Despite the excellent performance of HybrIK, the skeleton used in its estimation is composed of lines connecting various joints, derived from 2D HPE. For limbs considered rigid bodies, this skeleton-based model is an appropriate way to describe actions. However, the scenario becomes somewhat different in the case of the torso, which is a non-rigid body, as the skeleton becomes imperfect by disregarding or reducing bending information. While this model is effective for simple applications like action recognition, it becomes quite limited when it comes to motion analysis, spinal diagnostics, or other scenarios requiring accurate torso information.

The challenge in this last stage thus lay in representing the curvature of the spinal column. Some datasets add joints in the chest and abdomen to express body flexion. It's easy to consider adding multiple key points between the neck and the hip, so that the more of these points there are, the more similar the line connecting them becomes to the spinal curve. However, this ends up demanding more computational effort from the network. This is not only due to estimating a large number of key points, but also because the proximity between them makes their learning more difficult.

The proposed hypothesis was then the curve fitting through interpolation, which essentially involves the process of determining a function that will take on a set of known values at certain points, referred to as interpolation nodes. To obtain these nodes, we relied on the three-dimensional mesh of the patient resulting from SMPL and HybrIK. This representation, containing approximately 6890 vertices, is capable of depicting the human body with great precision. In addition to its vertices, in order to construct and depict the three-dimensional mesh, knowledge of the faces formed by these vertices, typically triangular, is equally essential, as they compose its surface.

Following the medical method for measuring the human spinal column, wherein the midline of the back can be approximated as the spine [70], the midline of the back surface of the mesh was taken into account. Thus, in an initial step, and to verify the accuracy of the mesh obtained by HybrIK to a certain extent, a Blender add-on was utilized, allowing the visualization of the patient's mesh, and the information pertaining to vertices and faces was stored in an object file. However, to make the INPACT platform as autonomous as possible, this task was eventually automated with a few lines of code. The object file created was then updated with new data at each execution frame of the exercise. This enabled, in the subsequent stage, the manual extraction of around 36 vertices from the central line of the back using a Python package known as Meshio, as illustrated in Figure 5.2. More importantly, it confirmed that this set of points consistently maintained the same index regardless of the exercise or position assumed by the patient. This greatly facilitated the subsequent interpolation process.

In this new interpolation task, we ended up utilizing a second Python library, SciPy, which stands out for containing several essential tools in solving mathematical, scientific, and engineering problems. This includes not only a wide range of optimization, integration, and differential equation algorithms, but also a diverse set of interpolation functions, which proved to be quite useful. By taking these functions and conducting some tests, it was possible to conclude that the two best functions for the given problem would be **interp1d** and **pchip_interpolate**. It is important to mention that the term PCHIP, Piecewise Cubic Hermite Interpolating Polynomial, arises from the fact that this second interpolator uses monotonic cubic splines to determine the values of the new points. Both interpolators receive as input the matrices of points x and y from the vertices extracted in the previous step. This detail is crucial because it requires that the first input matrix of x-coordinate points be properly ordered. That being said, it's easy to understand that for a certain set of exercises, such as the one in Figure 5.5, some of the points on the lower back of the patient will increase and decrease the value of their x-coordinate, leading to an unordered list. For this very reason, we opted for two interpolation functions, since the issue no longer occurs with **interp1d**.

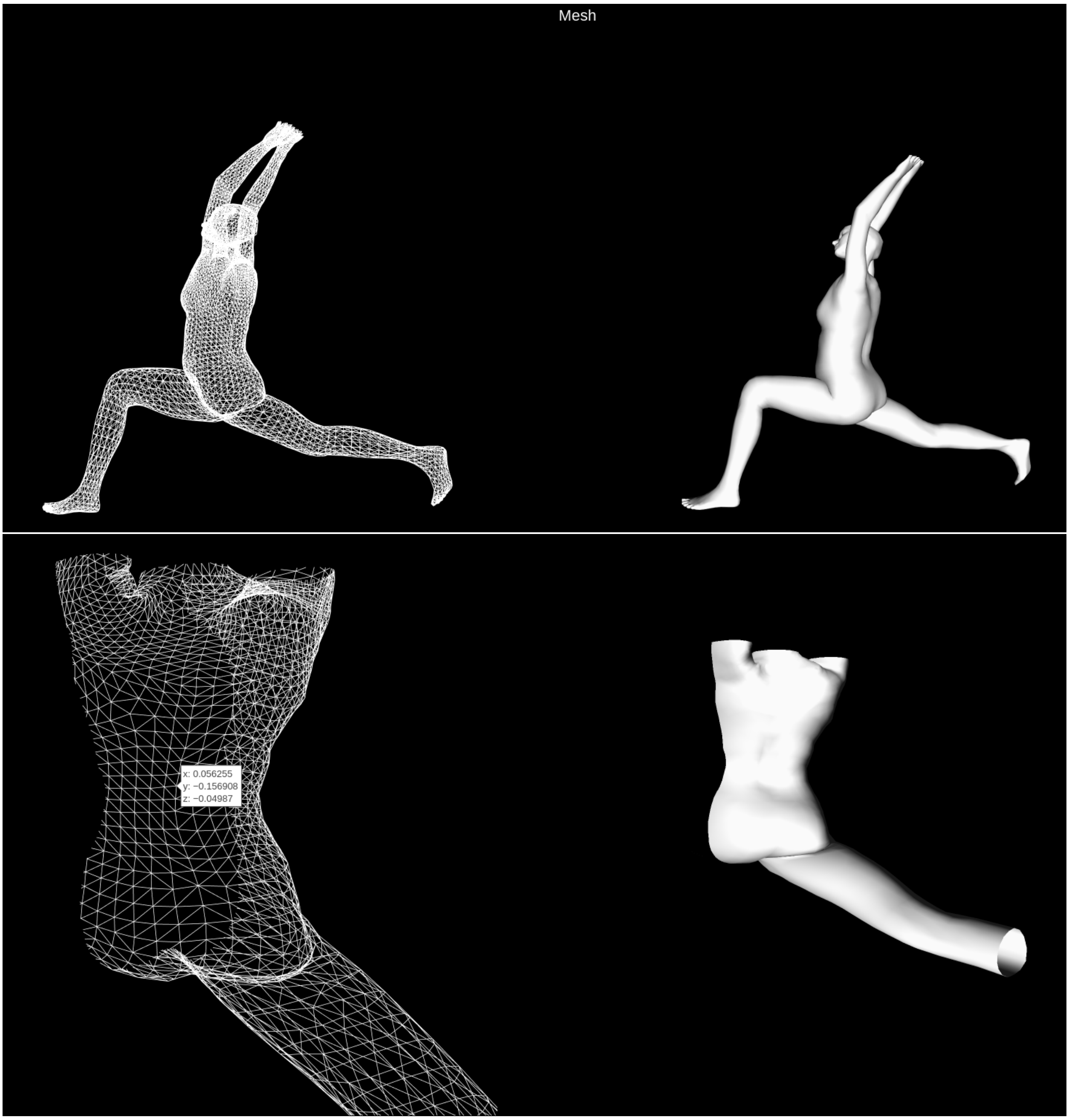


Figure 5.2: Visualization of the three-dimensional mesh using the Meshio library and extraction of the 36 vertices from the patient's back.

Finally, to complete the whole interpolation process and evaluate the accuracy of the obtained curve, we opted for a series of graphs where we superimposed the three-dimensional grids of each repetition estimated by HybrIK along with their respective curves. It's evident that the most effective way to visualize these results involves applying a rotation to both elements, such that they are positioned laterally, enabling the observed overlap. A prime example is the first squat exercise illustrated in Figure 5.3, where the individual is facing the camera. The approach employed entails the rotation of a plane formed by three points along the central line of the back (two at the ends and one central), effectively dividing the individual in half. The rotation applied to this imaginary plane is consequently mirrored onto the grid and its corresponding curve, as demonstrated in Figure 5.5. It is essential to note that although this evaluation is primarily visual, it was readily apparent that the derived curve is highly accurate and nearly coincides with the boundaries of the grid.

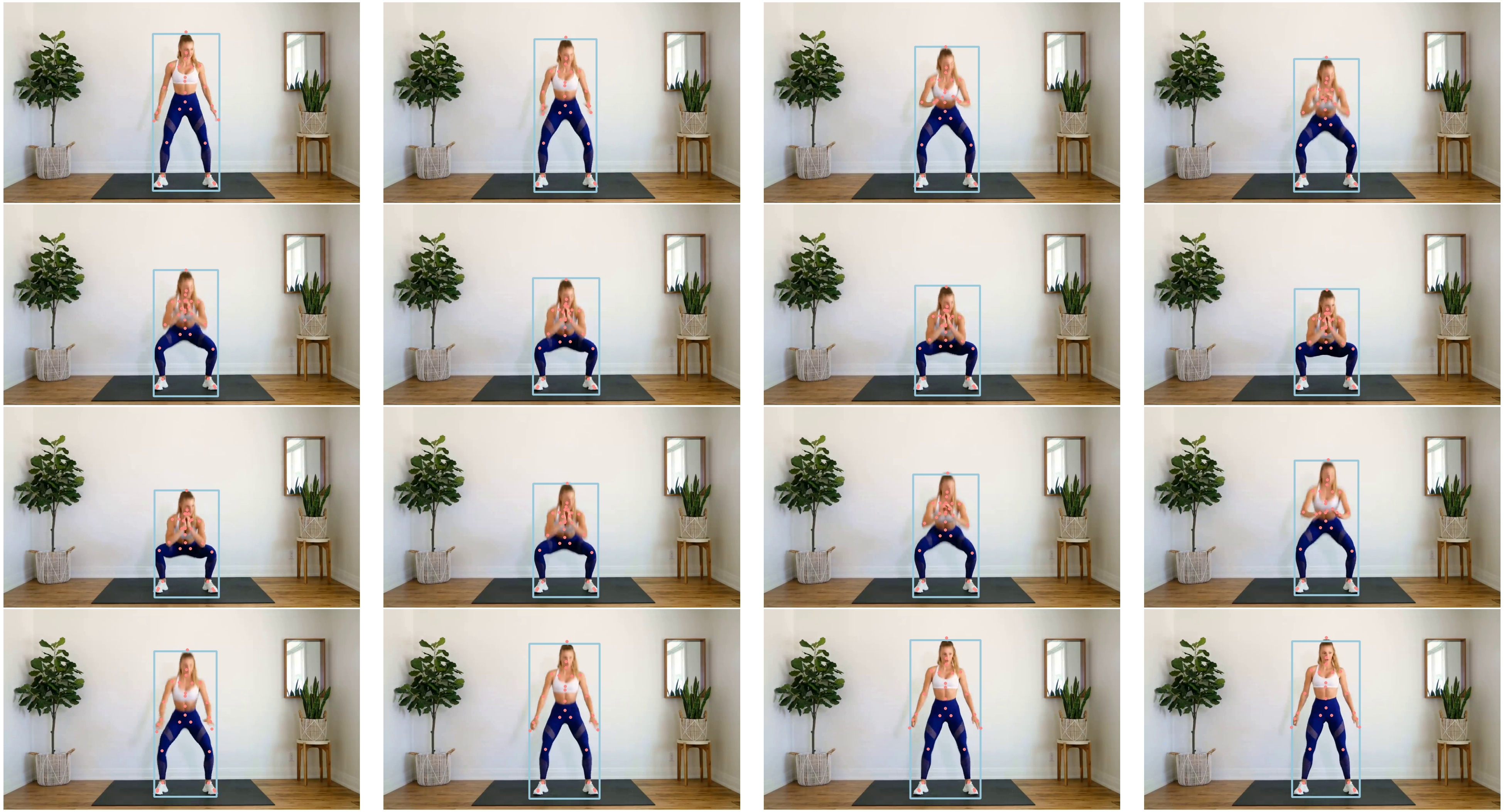


Figure 5.3: Sequence of 16 images in a squat exercise - Detection of the person and all his joints.

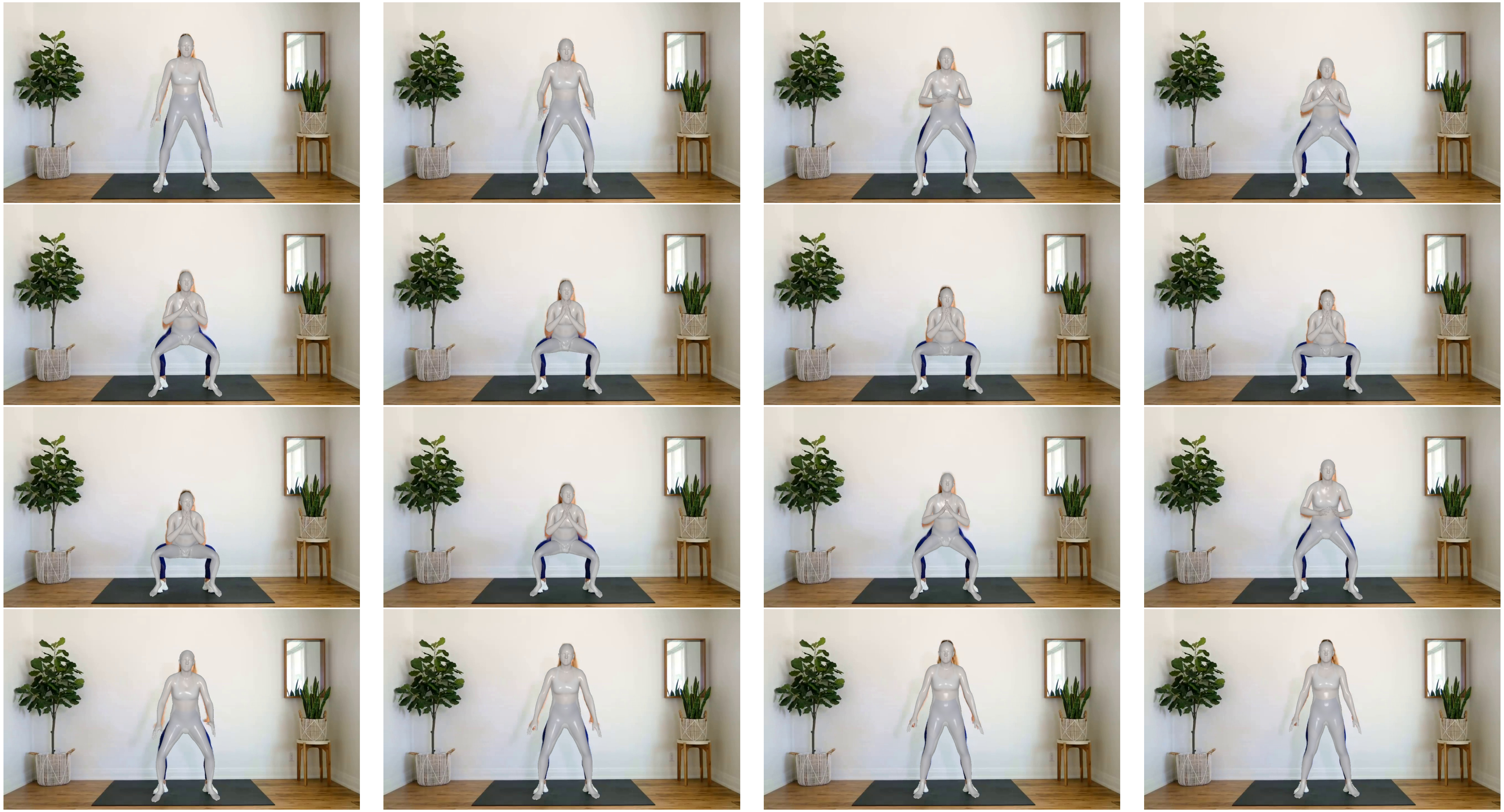


Figure 5.4: Sequence of 16 images in a squat exercise - Mesh prediction.

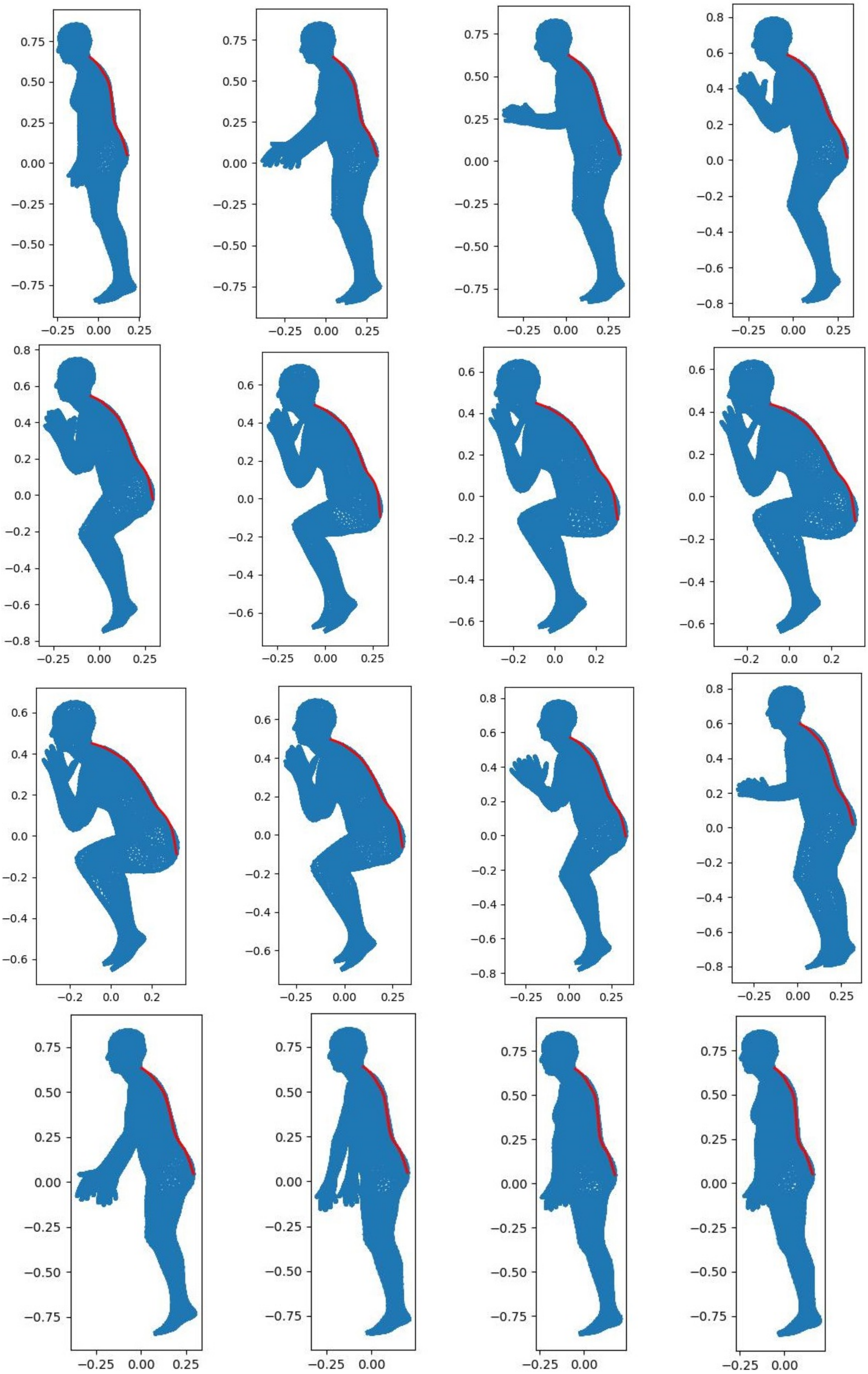


Figure 5.5: Sequence of 16 images in a squat exercise - Spine curvature segmentation.

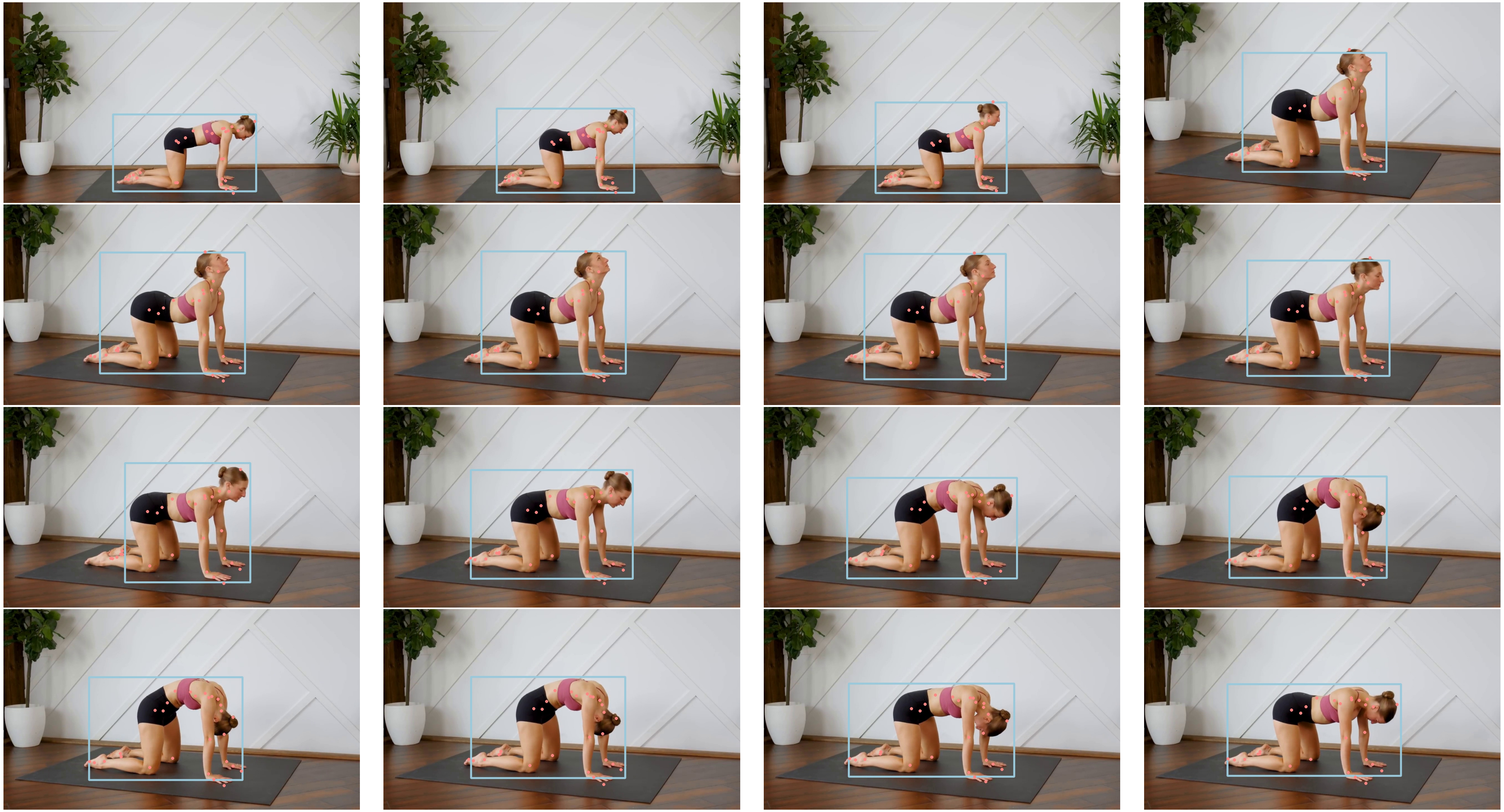


Figure 5.6: Sequence of 16 images in a cat stretch exercise - Detection of the person and all his joints.

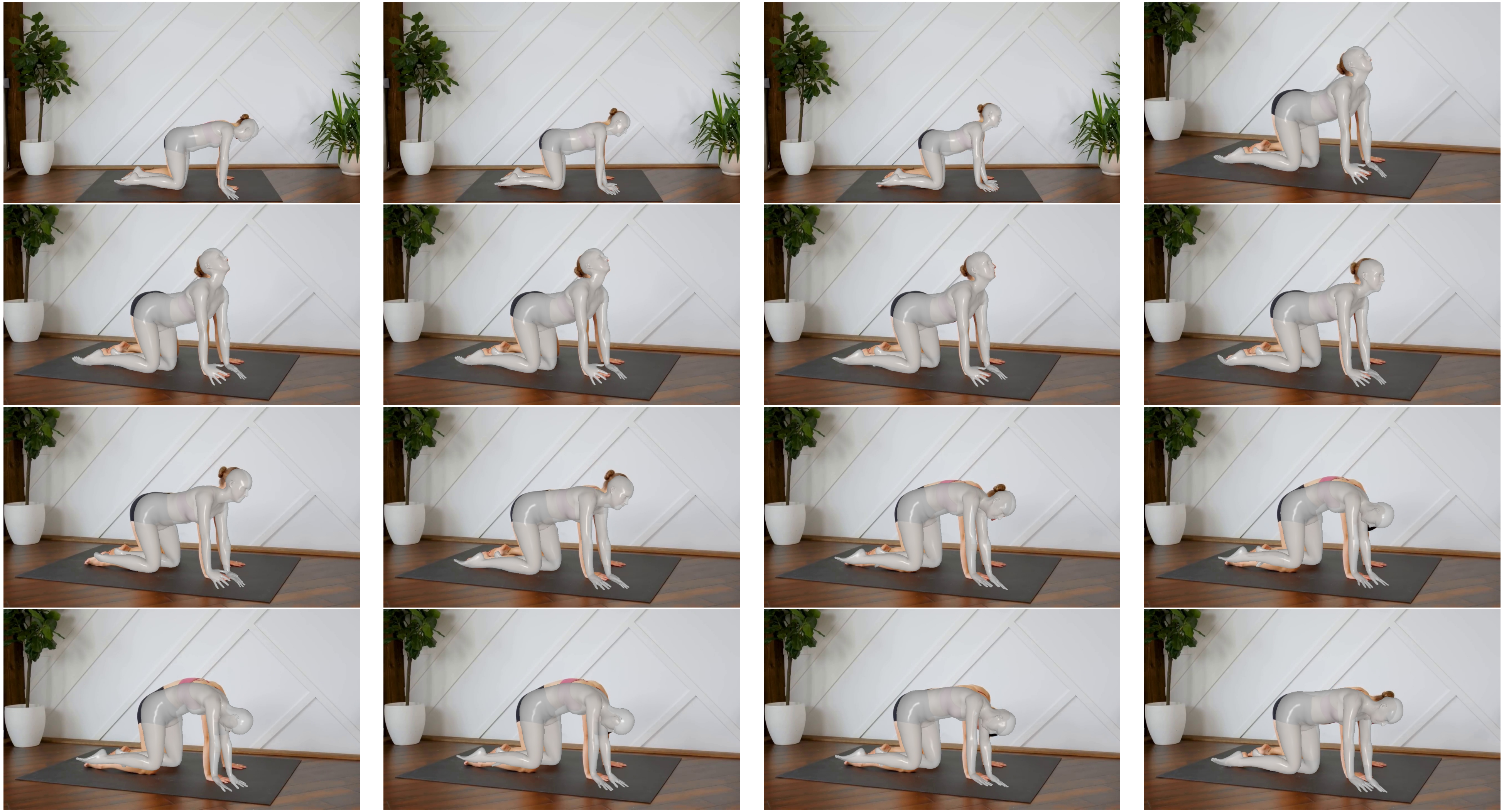
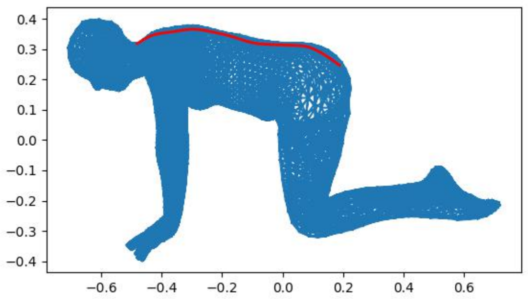
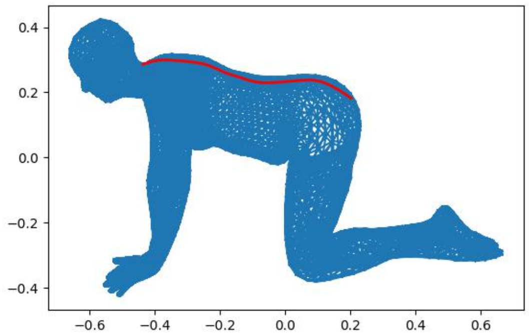
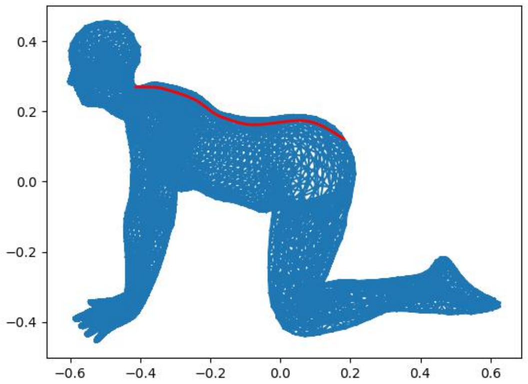
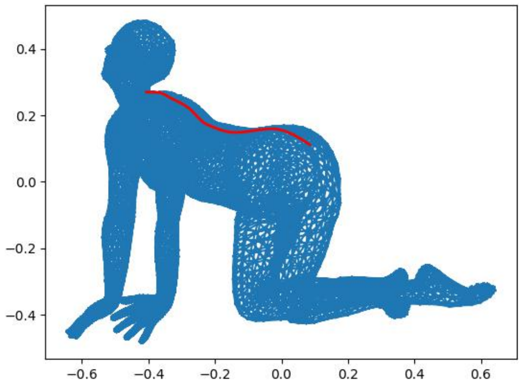


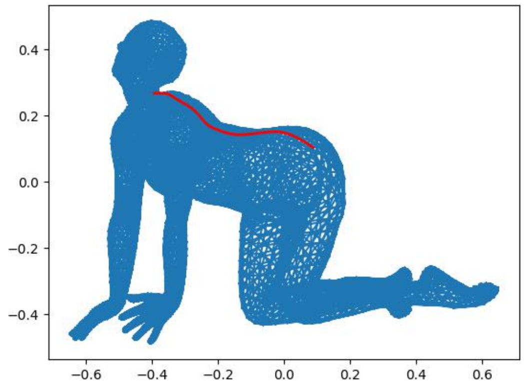
Figure 5.7: Sequence of 16 images in a cat stretch exercise - Mesh prediction.

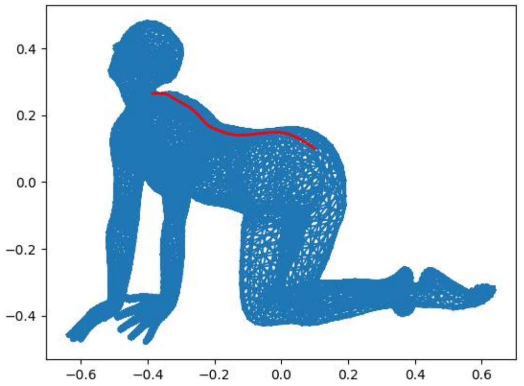

0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

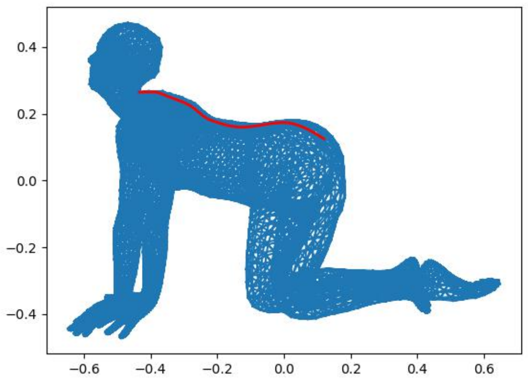

0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

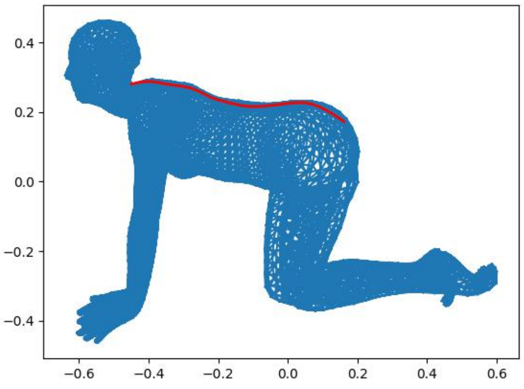

0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

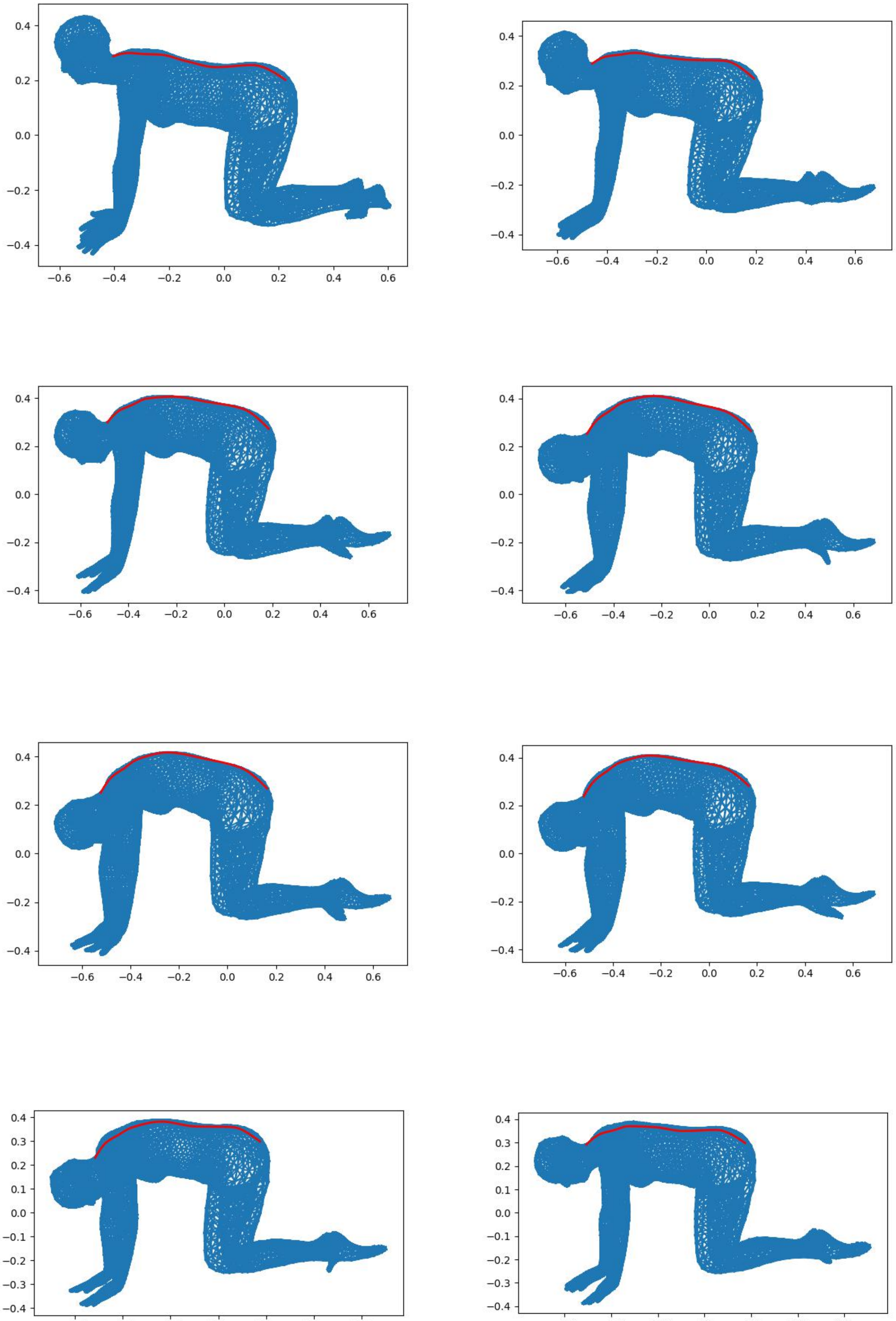


Figure 5.8: Sequence of 16 images in a cat stretch exercise - Spine curvature segmentation.

6 Conclusion

In conclusion, this thesis embarked on a journey to explore the intersection of telerehabilitation, back pain management, human pose and shape estimation, and the innovative use of the HybrIK method. Through meticulous research, experimentation, and analysis, we have successfully achieved all the objectives set forth in this work.

By integrating the HybrIK method into our study, we harnessed its power to accurately estimate human pose and shape, demonstrating its effectiveness in virtual environments for rehabilitation purposes. This novel approach has paved the way for more accessible and engaging telerehabilitation programs.

Moreover, we introduced a groundbreaking application of HybrIK by utilizing the vertices of virtual meshes to segment and analyse spine curvature. This innovation not only enhances our understanding of the biomechanics of the spine but also provides valuable insights for personalized rehabilitation strategies, tailored to individual needs.

As we conclude this thesis, we look forward to a future where telerehabilitation, informed by advancements in human pose and shape estimation techniques like HybrIK, plays a central role in mitigating the burden of back pain and improving the lives of countless individuals. This thesis is a testament to our commitment to innovation and our dedication to the betterment of healthcare for all.

7 Future work

The work carried out and presented always took into consideration the guidelines provided throughout the semester and was capable of producing the intended results. However, as not all systems can be perfect, there are always some aspects that require attention and improvement. One of these aspects includes the exercises from the previous examples. It is easily observed and understood that these exercises involve relatively simple movements and are easily estimated by HybrIK. However, in other cases, such as certain exercises for more complex rehabilitation that involve, for example, the occlusion and abnormal positioning of certain limbs, it was noticed that the model returned a highly deformed three-dimensional mesh of the individual. In this regard, we believe that the first course of action for future work would involve training the model with a specific rehabilitation dataset, such as Fit3D, produced by the same researchers from Human3.6. Apart from a variety of 37 exercises and repetitions, it contains over three million images and configurations of human motion capture, thus encompassing all major muscle groups. This would make the model specific to the field of telerehabilitation and enable a much more effective and accurate estimation through error correction.

The second topic, although not affecting the functioning of the model, is equally important. As can be observed in Figure 7.1, the patient's hands and feet exhibit unusual behaviour, and their face does not display any type of expression. This is due to the volumetric model used by SMPL and HybrIK, which does not include any type of joints in these regions. However, the HybrIK model has recently been updated with SMPL-X, an improved model of SMPL that allows for a much more realistic representation of the individual, including facial expressions and articulated hands. Thus, the work would involve using or training this new model, HybrIK-X, and verifying if its results show promise.

Lastly, as mentioned earlier, this dissertation was conducted within the scope of the INPACT project and involved the collaboration of a small group of students and professors. This encompassed not only the training and implementation of ML algorithms but also the creation of an interactive application for the user. Therefore, in order to obtain a functional final product that may potentially be produced on a large scale, there is the need to integrate the different parts.

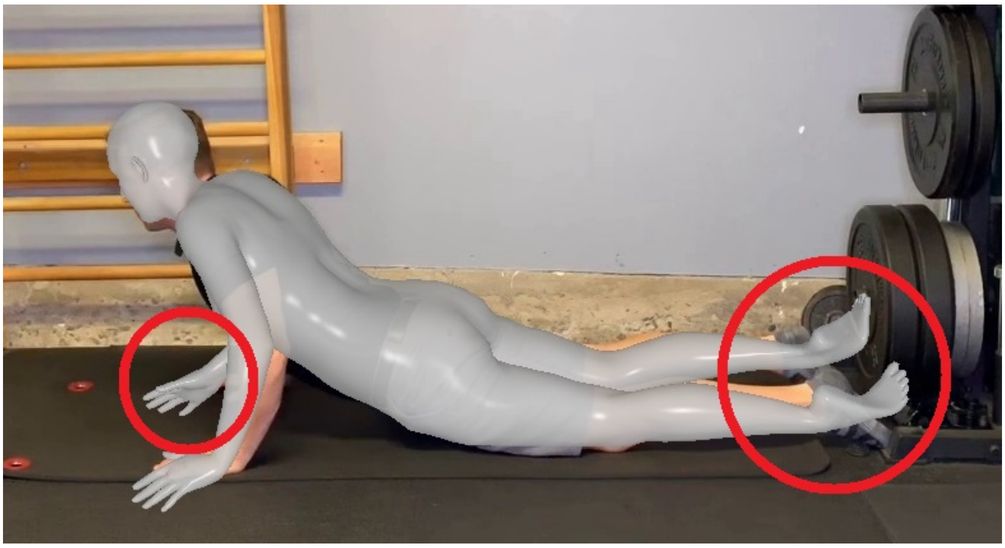


Figure 7.1: Illustrative example of an exercise in which the hands and feet of the three-dimensional model of the patient exhibit irregular behavior.

Bibliography

[1] S. Clark and R. Horton, "Low back pain: A major global challenge", *The Lancet*, vol. 391, no. 10137, p. 2302, 2018, ISSN: 0140-6736. DOI: `https://doi.org/10.1016/S0140-6736(18)30725-6`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0140673618307256`.

[2] R. F., "Applying vision-based pose estimation in a telerehabilitation application", 2021.

[3] A. Peretti., "Telerehabilitation: Review of the state-of-the-art and areas of application.", *JMIR Rehabil Assist Technol*, vol. 4, 2017.

[4] T. G. Russell., "Telerehabilitation: A coming of age", *Australian Journal of Physiotherapy.*, vol. 55, 2005.

[5] P. Amorim, J. Paulo, P. A. Silva, P. Peixoto, M. Castelo-Branco, and H. Martins, "Machine learning applied to low back pain rehabilitation – a systematic review", *International Journal of Digital Health*, vol. 1, p. 10, Apr. 2021. DOI: `10.29337/ijdh.34`.

[6] J. Wang, S. Tan, X. Zhen, *et al.*, "Deep 3d human pose estimation: A review", *Computer Vision and Image Understanding*, vol. 210, p. 103 225, 2021, ISSN: 1077-3142. DOI: `https://doi.org/10.1016/j.cviu.2021.103225`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1077314221000692`.

[7] K. Bartol, D. Bojanić, T. Petković, N. D'Apuzzo, and T. Pribanic, "A review of 3d human pose estimation from 2d images", Nov. 2020. DOI: `10.15221/20.29`.

[8] L. Chen, H. Ai, R. Chen, Z. Zhuang, and S. Liu, *Cross-view tracking for multi-human 3d pose estimation at over 100 fps*, 2021. arXiv: `2003.03972 [cs.CV]`.

[9] H. Qi, Y. Xu, T. Yuan, T. Wu, and S.-C. Zhu, *Scene-centric joint parsing of cross-view videos*, 2018. arXiv: `1709.05436 [cs.CV]`.

[10] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, *Fast and robust multi-person 3d pose estimation from multiple views*, 2019. arXiv: `1901.04111 [cs.CV]`.

[11] F. Galbusera., "Artificial intelligence and machine learning in spine research", 2019.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, *Deep high-resolution representation learning for human pose estimation*, 2019. arXiv: `1902.09212 [cs.CV]`.

[14] C. Zheng, W. Wu, T. Yang, *et al.*, "Deep learning-based human pose estimation: A survey", Dec. 2020.

[15] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods", *Computer Vision and Image Understanding*, vol. 192, p. 102 897, Mar. 2020. DOI: `10.1016/j.cviu.2019.102897`. [Online]. Available: `https://doi.org/10.1016%5C%2Fj.cviu.2019.102897`.

[16] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: Shape completion and animation of people", in *ACM SIGGRAPH 2005 Papers*, ser. SIGGRAPH '05, Los Angeles, California: Association for Computing Machinery, 2005, pp. 408–416, ISBN: 9781450378253. DOI: `10.1145/1186822.1073207`. [Online]. Available: `https://doi.org/10.1145/1186822.1073207`.

[17] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker, "Detailed human shape and pose from images", in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8. DOI: `10.1109/CVPR.2007.383340`.

[18] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, "Dyna: A model of dynamic human shape in motion", *ACM Trans. Graph.*, vol. 34, no. 4, Jul. 2015, ISSN: 0730-0301. DOI: `10.1145/2766993`. [Online]. Available: `https://doi.org/10.1145/2766993`.

[19] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model", *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, 248:1–248:16, Oct. 2015.

[20] S. Ren, K. He, R. Girshick, and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, 2016. arXiv: `1506.01497 [cs.CV]`.

[21] A. S. Micilotta, E.-J. Ong, and R. Bowden, "Real-time upper body detection and 3d pose estimation in monoscopic images", in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 139–150, ISBN: 978-3-540-33837-6.

[22] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks", in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2014. DOI: 10.1109/cvpr.2014.214. [Online]. Available: https://doi.org/10.1109%5C%2Fcvpr.2014.214.

[23] C. Szegedy, W. Liu, Y. Jia, *et al.*, *Going deeper with convolutions*, 2014. arXiv: 1409.4842 [cs.CV].

[24] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, *Human pose estimation with iterative error feedback*, 2016. arXiv: 1507.06550 [cs.CV].

[25] S. Li, Z.-Q. Liu, and A. B. Chan, *Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network*, 2014. arXiv: 1406.3474 [cs.CV].

[26] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, *Convolutional pose machines*, 2016. arXiv: 1602.00134 [cs.CV].

[27] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation", in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 197–214, ISBN: 978-3-030-01219-9.

[28] W. Tang and Y. Wu, "Does learning specific features for related parts help human pose estimation?", in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1107–1116. DOI: 10.1109/CVPR.2019.00120.

[29] B. Xiao, H. Wu, and Y. Wei, *Simple baselines for human pose estimation and tracking*, 2018. arXiv: 1804.06208 [cs.CV].

[30] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, *Rmpe: Regional multi-person pose estimation*, 2018. arXiv: 1612.00137 [cs.CV].

[31] L. Pishchulin, E. Insafutdinov, S. Tang, *et al.*, *Deepcut: Joint subset partition and labeling for multi person pose estimation*, 2016. arXiv: 1511.06645 [cs.CV].

[32] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, *Deepercut: A deeper, stronger, and faster multi-person pose estimation model*, 2016. arXiv: 1605.03170 [cs.CV].

[33] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, *Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model*, 2018. arXiv: `1803.08225 [cs.CV]`.

[34] M. Kocabas, S. Karagoz, and E. Akbas, *Multiposenet: Fast multi-person pose estimation using pose residual network*, 2018. arXiv: `1807.04067 [cs.CV]`.

[35] Z. Li, "3d human pose and shape estimation based on parametric model and deep learning", English, Ph.D. dissertation, 2021, ISBN: 978-91-7895-787-3.

[36] A. Balan, L. Sigal, M. J. Black, J. Davis, and H. Haussecker, "Detailed human shape and pose from images", in *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR*, Minneapolis, Jun. 2007, pp. 1–8.

[37] P. Guan, A. Weiss, A. Balan, and M. J. Black, "Estimating human shape and pose from a single image", in *Int. Conf. on Computer Vision, ICCV*, 2009, pp. 1381–1388.

[38] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular RGB-D sequences", in *International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2300–2308.

[39] L. Pishchulin, E. Insafutdinov, S. Tang, *et al.*, *Deepcut: Joint subset partition and labeling for multi person pose estimation*, 2016. arXiv: `1511.06645 [cs.CV]`.

[40] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, *Video based reconstruction of 3d people models*, 2018. arXiv: `1803.04758 [cs.CV]`.

[41] W. Xu, A. Chatterjee, M. Zollhöfer, *et al.*, *Monoperfcap: Human performance capture from monocular video*, 2018. arXiv: `1708.02136 [cs.CV]`.

[42] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross, "Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks", in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 108–117. DOI: `10.1109/3DV.2016.19`.

[43] I. B. Vince Tan and R. Cipolla, "Indirect deep structured learning for 3d human body shape and pose prediction", in *Proceedings of the British Machine Vision Conference (BMVC)*, G. B. Tae-Kyun Kim Stefanos Zafeiriou and K. Mikolajczyk, Eds., BMVA Press, Sep. 2017, pp. 15.1–15.11, ISBN: 1-901725-60-X. DOI: `10.5244/C.31.15`. [Online]. Available: `https://dx.doi.org/10.5244/C.31.15`.

[44] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, *Unite the people: Closing the loop between 3d and 2d human representations*, 2017. arXiv: `1701.02468 [cs.CV]`.

[45] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image", in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science, Springer International Publishing, Oct. 2016.

[46] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, *End-to-end recovery of human shape and pose*, 2018. arXiv: `1712.06584 [cs.CV]`.

[47] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, *Learning to estimate 3d human pose and shape from a single color image*, 2018. arXiv: `1805.04092 [cs.CV]`.

[48] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, *Learning to reconstruct 3d human pose and shape via model-fitting in the loop*, 2019. arXiv: `1909.12828 [cs.CV]`.

[49] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network", in *Asian Conference on Computer Vision*, 2014.

[50] S. Li, W. Zhang, and A. B. Chan, *Maximum-margin structured learning with deep networks for 3d human pose estimation*, 2015. arXiv: `1508.06708 [cs.CV]`.

[51] X. Sun, J. Shang, S. Liang, and Y. Wei, *Compositional human pose regression*, 2017. arXiv: `1704.00159 [cs.CV]`.

[52] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, *Coarse-to-fine volumetric prediction for single-image 3d human pose*, 2017. arXiv: `1611.07828 [cs.CV]`.

[53] C.-H. Chen and D. Ramanan, *3d human pose estimation = 2d pose estimation + matching*, 2017. arXiv: `1612.06524 [cs.CV]`.

[54] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation", in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2659–2668. DOI: `10.1109/ICCV.2017.288`.

[55] K. Zhou, X. Han, N. Jiang, K. Jia, and J. Lu, *Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation*, 2019. arXiv: `1910.12032 [cs.CV]`.

[56] F. Moreno-Noguer, *3d human pose estimation from a single image via distance matrix regression*, 2016. arXiv: `1611.09010 [cs.CV]`.

[57] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, *Monocular 3d human pose estimation by generation and ordinal ranking*, 2019. arXiv: 1904.01324 [cs.CV].

[58] C. Li and G. H. Lee, *Generating multiple hypotheses for 3d human pose estimation with mixture density network*, 2019. arXiv: 1904.05547 [cs.CV].

[59] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll, and C. Theobalt, *In the wild human pose estimation using explicit 2d features and intermediate 3d representations*, 2019. arXiv: 1904.03289 [cs.CV].

[60] M. Ropars, A. Crétual, H. Thomazeau, R. Kaila, and I. Bonan, "Volumetric definition of shoulder range of motion and its correlation with clinical signs of shoulder hyperlaxity. A motion capture study.", *Journal of Shoulder and Elbow Surgery*, vol. 24, no. 2, pp. 310–316, 2015. DOI: 10.1016/j.jse.2014.06.040. [Online]. Available: https://inria.hal.science/hal-01058988.

[61] R. Bednarski and A. Bielak, "Use of motion capture in assisted of knee ligament injury diagnosis", Jan. 2018.

[62] N. Kleanthous, A. J. Hussain, W. Khan, and P. Liatsis, "A new machine learning based approach to predict freezing of gait", *Pattern Recognition Letters*, vol. 140, pp. 119–126, 2020, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2020.09.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520303524.

[63] M. Madadi, H. Bertiche, and S. Escalera, "Smplr: Deep learning based smpl reverse for 3d human pose and shape recovery", *Pattern Recognition*, vol. 106, p. 107 472, 2020, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2020.107472. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320302752.

[64] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383–3393.

[65] 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, Jul. 2014.

[66] D. Mehta, H. Rhodin, D. Casas, *et al.*, "Monocular 3d human pose estimation in the wild using improved cnn supervision", in *3D Vision (3DV), 2017 Fifth International Conference on*, IEEE, 2017. DOI: 10.1109/3dv.2017.00064. [Online]. Available: http://gvv.mpi-inf.mpg.de/3dhp_dataset.

[67] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, *Microsoft coco: Common objects in context*, 2015. arXiv: 1405.0312 [cs.CV].

[68] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera", in *European Conference on Computer Vision (ECCV)*, Sep. 2018.

[69] G. Bradski, "The opencv library", *Dr. Dobb's Journal of Software Tools*, 2000.

[70] C. Antonya, S. Butnariu, and C. Pozna, "Real-time representation of the human spine with absolute orientation sensors", in *2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2016, pp. 1–6. DOI: 10.1109/ICARCV.2016.7838745.